APLICACIÓN DE LA MINERÍA DE DATOS PARA LA DETECCIÓN DE FACTORES ASOCIADOS A LA DESERCIÓN ESTUDIANTIL EN LOS PROGRAMAS PROFESIONALES DE LAS CIENCIAS NATURALES Y CIENCIAS SOCIALES DE LA UNIVERSIDAD DE NARIÑO SEDE PASTO

EDUAR ALIRIO ERAZO ORTIZ
SERGIO ALDEMAR MORA PORTILLA

UNIVERSIDAD TECNOLÓGICA DE PEREIRA
FACULTAD DE CIENCIAS EMPRESARIALES
MAESTRÍA EN INVESTIGACIÓN OPERATIVA Y ESTADÍSTICA
SAN JUAN DE PASTO
2018

APLICACIÓN DE LA MINERÍA DE DATOS PARA LA DETECCIÓN DE FACTORES ASOCIADOS A LA DESERCIÓN ESTUDIANTIL EN LOS PROGRAMAS PROFESIONALES DE LAS CIENCIAS NATURALES Y CIENCIAS SOCIALES DE LA UNIVERSIDAD DE NARIÑO SEDE PASTO

EDUAR ALIRIO ERAZO ORTIZ
SERGIO ALDEMAR MORA PORTILLA

TRABAJO DE GRADO PRESENTADO COMO REQUISITO PARA OPTAR EL TÍTULO DE MAGISTER EN INVESTIGACIÓN OPERATIVA Y ESTADÍSTICA

DIRECTOR:
PhD. RICARDO TIMARÁN PEREIRA

UNIVERSIDAD TECNOLÓGICA DE PEREIRA
FACULTAD DE CIENCIAS EMPRESARIALES
MAESTRÍA EN INVESTIGACIÓN OPERATIVA Y ESTADÍSTICA
SAN JUAN DE PASTO
2018

AGRADECIMIENTOS

Al PhD. Ricardo Timarán Pereira por su acompañamiento, sus concejos, sugerencias y valiosa colaboración.

A la Universidad de Nariño por su gestión ante la UTP para que se lleve a cabo la Maestría en Investigación Operativa y Estadística. En especial, a su Facultad de Ciencias Exactas y Naturales que a través de su director; Mg. Hernán García se fue posible optar a una beca para que uno de los integrantes curse esta maestría.

Al director de la Maestría en Investigación Operativa y Estadística de la Universidad Tecnológica de Pereira PhD. José Soto Mejía por su decisión de hacer viable el desarrollo de la misma en la ciudad de Pasto, buscando siempre para la formación de sus estudiantes a los profesionales más idóneos.

RESUMEN

En este documento se presentan los resultados de la investigación que tuvo como objetivo descubrir los factores asociados a la deserción estudiantil en los programas de pregrado de la Universidad de Nariño sede Pasto, tanto de las Ciencias Naturales como de las Ciencias Sociales y Humanas, a partir de los datos socioeconómicos, académicos, institucionales y de admisión registrados en las bases de datos de los estudiantes de dicha universidad, aplicando técnicas de Análisis Estadístico y de Minería de Datos. Se utilizó como metodología CRISP-DM (*Cross-Industry Standard Process for Data Mining*), uno de los modelos más utilizados en proyectos de minería de datos.

Se construyeron dos repositorios de datos: uno, con la información de los estudiantes que ingresaron en las cohortes 2008A hasta el periodo 2011B con una ventana de observación de 6 años; y otro, con la información histórica de los estudiantes que ingresaron desde el 2006 hasta el 2015. Ambos contienen información socioeconómica, académica e institucional de los estudiantes de los diferentes programas de pregrado de la Universidad de Nariño Sede Pasto, agrupados en Ciencias Naturales y Ciencias Sociales y Humanas. A estos repositorios se le aplicaron técnicas de limpieza y transformación.

Para la fase de minería de datos se empleó la herramienta de software libre weka3.8.2 con la técnica de clasificación Bayes con el algoritmo BayesNet y árboles de decisión con los algoritmos J48, Random Forest y LMT ***(Logistic Model Tree)***, desatacándose J48 como el modelo con mayor exactitud y precisión. El tener un promedio de notas bajo, menor que 3 y el número de materias perdidas (mínimo 1), son el patrón general de la deserción de los estudiantes de la Universidad de Nariño sede Pasto.

El conocimiento descubierto ayudará a soportar de manera efectiva la toma de decisiones de las directivas académicas de la Universidad de Nariño para la formulación de planes y programas que ayuden a minimizar la deserción y conlleven al mejoramiento de la calidad educativa en la institución.

ABSTRACT

This document presents the results of the research that aimed to discover the factors associated with student desertion in undergraduate programs, both in the Natural Sciences and in the Social and Human Sciences, based on socio-economic, academic data, institutional and admissions registered in the databases of the students of the University of Nariño headquarters Pasto, applying techniques of Statistical Analysis and Data Mining. It was used as CRISP-DM methodology (Cross-Industry Standard Process for Data Mining), one of the most used models in data mining projects.

Two repositories of data were constructed: one, with the information of the students who entered the cohorts 2008A until the period 2011B with a window of observation of 6 years; and another, with the historical information of the students who entered from 2006 to 2015. Both contain socio-economic, academic and institutional information of the students of the different undergraduate programs of the University of Nariño headquarters Pasto, grouped in Natural Sciences and Sciences Social and Human. Recycling and transformation techniques were applied to these repositories.

For the data mining phase we used the free software tool weka3.8.2 with the Bayes classification technique with the BayesNet algorithm and decision trees with the J48, Random Forest and LMT(Logistic Model Tree) algorithms, untied J48 as the model with greater accuracy and precision. The average of low grades, less than 3 and the number of lost subjects (minimum 1), is the general pattern of the dropout of the students of the University of Nariño, Pasto.

The knowledge discovered will help to effectively support the decision making of the academic directives of the University of Nariño for the formulation of plans and programs that help to minimize the desertion and lead to the improvement of the educational quality in the institution

TABLA DE CONTENIDO

LISTA DE TABLAS

LISTA DE FIGURAS

1. ASPECTOS PRELIMINARES

1.1. Introducción

La deserción se convirtió en un problema cuya preocupación es del orden internacional, su solución interesa en la actualidad tanto a los gobiernos como a las directivas de las instituciones educativas; cada estudiante que abandona su programa de estudios implica una pérdida de esfuerzos personales, recursos familiares y estatales, en este sentido deja de ser un problema meramente individual y pasa a ser un problema que compromete la política educativa nacional.

En los países como Colombia, donde acceder a la educación pública superior es un privilegio, cada estudiante que deserta se constituye en una afrenta que viene a lesionar los intereses de amplios sectores de la sociedad: detrás de un cupo hay centenas de aspirantes que no lograron ingresar al programa.

En el Departamento de Nariño, la mayor parte de estudiantes universitarios se encuentran en la Universidad de Nariño, la cual, es una institución pública Acreditada de Alta Calidad por el Ministerio de Educación Nacional, y reconocida por su nivel educativo y prestigio de sus egresados. Desafortunadamente, algunos estudiantes, por diferentes razones, no logran culminar sus estudios de pregrado, generando el interrogante acerca de cuáles pueden ser las causas que conllevan a la deserción estudiantil universitaria.

En este proyecto se descubren factores asociados a la deserción estudiantil en los distintos programas profesionales de las Ciencias Naturales y las Ciencias Sociales y Humanas de la Universidad de Nariño Sede Pasto, utilizando técnicas estadísticas y de minería de datos en las variables socioeconómicas, académicas, institucionales y de admisión, que se encuentran almacenadas en las bases de datos de la Universidad de Nariño. Para tal fin, se construye dos repositorios de datos; uno, con la información de los estudiantes que ingresaron en las cohortes 2008A hasta el periodo 2011B con una ventana de observación de 6 años; y otro, con la información histórica de los estudiantes que ingresaron desde el año 2006 hasta el 2015. En cada uno de ellos se determina el índice de deserción estudiantil. Una vez procesados los datos se aplican técnicas estadísticas y de minería de datos para el descubrimiento de patrones determinantes de la deserción estudiantil en la Universidad de Nariño. El conocimiento adquirido ayudará a soportar de manera efectiva la toma de decisiones por parte de las directivas académicas de la Universidad de Nariño para la formulación de planes y programas que ayuden a minimizar la tasa de deserción estudiantil.

1.2. Planteamiento Del Problema

1.2.1. Descripción del problema

Para el Ministerio de Educación Nacional, en Latinoamérica, la educación superior presenta altas tasas de deserción estudiantil, básicamente en los primeros semestres, lo cual, genera efectos de tipo económico, académico y social, tanto para las Instituciones de Educación Superior (IES) como para el estudiante, la región y el país [16]. Sin embargo, a éste fenómeno no se le ha prestado la atención necesaria y mucho menos existen políticas formales para afrontarlo. Así mismo, en [16] se concluye que en Colombia, por lo menos el 52% de los estudiantes que empiezan una carrera universitaria no la concluyen. Además, se afirma que de las promociones de estudiantes que iniciaron estudios entre 1999 y el 2004, en promedio el 48% finalizaron sus estudios. Es decir, de cada dos estudiantes que se matriculan en un programa de pregrado, sólo uno culmina su carrera. También señala, que el 39,52% de quienes abandonan sus estudios lo hacen por razones económicas.

En países como Colombia, donde acceder a la educación pública superior es un privilegio, cada estudiante que deserta se constituye en una afrenta que viene a lesionar los intereses de amplios sectores de la sociedad, dado que, detrás de un cupo hay centenas de aspirantes que no lograron ingresar al programa [39].

Para el año 2005, en Colombia, la cobertura en la Educación Superior fue del 21,5% de la población escolar [15], y de esta, más de la mitad de los estudiantes matriculados abandonan sus estudios sin obtener un título profesional, especialmente, durante los primeros semestres. La mayoría de las Instituciones de Educación Superior (IES) colombianas han implementado mecanismos para disminuir estos índices, a saber: incremento de los cupos universitarios, sistemas alternativos de financiación, flexibilidad académica, programas de seguimiento y apoyo a posibles desertores, mejoramiento del bienestar universitario, entre otros. Sin embargo, muchos de estos esfuerzos no han sido suficientes y el fenómeno continúa repitiéndose, tal como ocurre en la Universidad de Nariño, la cual, no ha logrado salir de ésta problemática.

Analizando los datos de ingresos de estudiantes en los programas de pregrado de la Universidad de Nariño, en el periodo comprendido entre 2010A y 2015A se matricularon 11.044 estudiantes, según datos proporcionados por la Oficina de Control y Registro Académico OCARA. De acuerdo con el Anuario Estadístico de la Universidad de Nariño 2010-2015, la tasa de deserción estudiantil en ese periodo fue de 41,66%. Esto significa que de 11.044 estudiantes que se matricularon en este periodo, 4601 estudiantes se retiraron, hecho que genera una pérdida de esfuerzos personales, recursos familiares, estatales y de la Universidad de Nariño.

En vista de la importancia de esta situación se plantea un interrogante: ¿Cuáles son los factores principales asociados a la deserción estudiantil en los programas de pregrado, tanto de las Ciencias Naturales como de las Ciencias Sociales y Humanas de la Universidad de Nariño sede Pasto?

En este proyecto se declara algunos aspectos asociados a la deserción estudiantil en los programas de pregrado clasificados en Ciencias Naturales o Ciencias Sociales y Humanas de la Universidad de Nariño Sede Pasto, utilizando técnicas estadísticas y de minería de datos, a partir del uso de aspectos socioeconómicos, académicos, institucionales y de admisión de los de estudiantes, con el fin de brindar información de calidad que permita soportar de manera efectiva la toma de decisiones de las directivas académicas de la Universidad, para la formulación de planes y programas que ayuden a minimizar la deserción y conlleven al mejoramiento de la calidad educativa en la institución.

1.2.2. Delimitación del problema

Esta investigación hizo uso de los datos socioeconómicos, académicos y de admisión de estudiantes de pregrado de los diferentes programas profesionales de la Universidad de Nariño Sede Pasto, pertenecientes a las cohortes 2006A hasta el 2011B con un periodo de observación de 6 años. Además, por otro lado se empleó la información histórica de los estudiantes que ingresaron desde el 2006 hasta el 2015[1]. Con lo cual, se determinaron tanto factores asociados a la deserción estudiantil como el índice de deserción (calculado como la razón entre el número de desertores y número de estudiantes) [28].

A nivel de Colombia, el Ministerio de Educación Nacional, quien es el veedor de los intereses de la educación del país, ha propuesto una definición conjugada de deserción basada en Tinto [33] y Giovagnoli [7], la cual, considera como “desertor”, aquel estudiante que, en el momento en que se observa, ha abandonado durante dos o más periodos consecutivos la institución o no registra graduación [14]. De ahí que la presente investigación ha tomado como punto de referencia ésta definición.

Mediante el sistema gestor de bases de datos (SGBD) postgresSQL que es un sistema de código abierto de fácil acceso y de bastante ayuda para consultas y organización de grandes bases de datos, se caracterizó a los estudiantes desertores en dos grupos, los pertenecientes a los programas de pregrado de las Ciencias Sociales y Humanas y estudiantes de las Ciencias Naturales, esto, debido a la diferencia en el currículo de cada conjunto de programas académicos. Entendiéndose como Ciencias Sociales y Humanas, a aquellos programas de pregrado de la Universidad de Nariño Sede Pasto que están adscritos a las Facultades de: Educación, Medicina, Derecho, Ciencias Humanas, Artes y Ciencias

[1] El periodo de observación para el repositorio Histórico se tomó hasta diciembre de 2017, dado que no importa si la cohorte ha terminado o no, mientras que, para el repositorio por cohortes la observación se realizó hasta junio de 2017, dado que interesa tener estudiantes graduados y analizar la deserción mediante la definición proporcionada por el MEN.

Económicas. En Ciencias Naturales se consideró los programas de las Facultades: Ciencias Exactas, Ciencias Agrícolas, Ingeniería, Ciencias Pecuarias e Ingeniería Agroindustrial.

La presente investigación adoptó el clasificador J48 basado en árboles de decisión, dado que es el más exitoso en cuanto a la predicción de una variable categórica, en comparación a otros modelos como BayesNet, Random Forest, y LMT, en [12],[17] y [25] se muestra que al igual que en nuestro caso j48 es el mejor modelo de predicción tanto para variables académicas de salud, industria etc. Así mismo, se hace uso de la herramienta de software libre Weka 3.8.2, que cuenta con el algoritmo J48, el cual es un instrumento de clasificación que implementa el algoritmo C4.5 (una mejora de id3) y se basa en la utilización del criterio del coeficiente de ganancia de información (information gain ratio). Este, incorpora una poda del árbol de clasificación como lo es el factor de confianza C (confidence level), que influye en el tamaño y capacidad de predicción. Cuanto más baja se haga esa probabilidad, más se exigirá para que la diferencia en los errores de predicción antes y después de podar sea más significativa. El valor por defecto de este factor es del 25%, y conforme va bajando este valor se permiten más operaciones de poda; por lo tanto, se puede llegar a árboles cada vez más pequeños [6]. Otra forma de variar el tamaño del árbol que utiliza este algoritmo es a través del parámetro M que especifica el mínimo número de instancias o registros por nodo del árbol; es menos importante puesto que depende del número absoluto de instancias en el conjunto de datos de partida [8].

1.2.3. Viabilidad de la investigación

Siendo la deserción estudiantil universitaria motivo de investigación de orden internacional, junto al auge que ha tomado el estudio de esta problemática en los últimos años, y en vista de que el Centro de Informática de la Universidad de Nariño Sede Pasto estuvo presta a facilitar la información requerida, fue viable llevar a cabo este proyecto.

1.2.4. Limitantes de la investigación

Una de las limitaciones de la investigación fue la modalidad de ingreso a la Universidad, la cual, está basada en las pruebas Saber 11°, que en los últimos años han tenido ciertas modificaciones, tales como: cambio en la forma de calificación (periodo 2010-2), eliminación o agregación asignaturas a evaluar. De ahí que para varios registros hay datos faltantes.

Otra limitante, es que no es posible garantizar la veracidad de los datos suministrados por los estudiantes al momento de inscripción y admisión. Con el propósito de obtener información más precisa fue necesario tomar los datos de las cohortes 2008A hasta 2011B, despreciando los registros de estudiantes admitidos entre 2006A y 2007B dado que éstos presentaron un alto porcentaje de datos nulos y faltantes.

1.3. Justificación

La deserción estudiantil es un problema que aqueja al Sistema de Educación Superior, cuya preocupación es de orden internacional, ya que, los cupos de admisión son limitados y en consecuencia, al desertar se le niega la posibilidad de estudio a otras personas que quizá desistan y se empleen en labores poco productivas, acarreando un aumento en el círculo de pobreza, además, se genera un desperdicio de recursos por parte de las instituciones, del estudiante y de sus familias.

La solución a éste fenómeno, en la actualidad, es motivo de interés tanto para los entes gubernamentales como para las directivas de las instituciones educativas, incluyendo la Universidad de Nariño. En consecuencia, se planteó este proyecto de investigación cuyo objetivo fue detectar factores asociados a la deserción estudiantil en los programas de pregrado, tanto de las Ciencias Naturales como de las Ciencias Sociales y Humanas de la Universidad de Nariño Sede Pasto, utilizando técnicas de análisis estadístico y de minería de datos para examinar este fenómeno y caracterizar a los estudiantes desertores, teniendo en cuenta aspectos como los socioeconómicos, académicos, institucionales y de admisión.

Con la realización de este estudio, las directivas académicas de la Universidad de Nariño contarán con el conocimiento necesario para soportar de manera efectiva la toma de decisiones en lo que se refiere a la formulación de planes y programas que ayuden a minimizar este fenómeno y aumentar la retención estudiantil, en beneficio de toda la comunidad académica y sociedad nariñense.

1.4. Objetivos Del Proyecto

1.4.1. Objetivo general

Descubrir factores asociados a la deserción estudiantil en los programas de pregrado, tanto de las Ciencias Naturales como de las Ciencias Sociales y Humanas, a partir de los datos socioeconómicos, académicos, institucionales y de admisión registrados en las bases de datos de los estudiantes de la Universidad de Nariño Sede Pasto, utilizando técnicas de Análisis Estadístico y de Minería de Datos, con el fin de disponer de información que permita soportar de manera efectiva la toma de decisiones de sus directivas académicas para la formulación de planes y programas que ayuden a minimizar la deserción estudiantil.

1.4.2. Objetivos específicos

Apropiar el conocimiento sobre deserción estudiantil por parte de los investigadores.

Identificar y seleccionar de las bases de datos internas y externas de la Universidad de Nariño, los datos socioeconómicos, académicos, institucionales y de admisión de los estudiantes de pregrado de la Sede Pasto.

Construir dos repositorios de datos con los aspectos socioeconómicos, académicos, institucionales y de admisión de los estudiantes. Uno por cohortes y otro histórico.

Aplicar técnicas de pre-procesamiento y transformación de datos a los repositorios con el fin de obtener registros limpios, correctos, consistentes y categorizados.

Aplicar las técnicas estadísticas y de minería de datos más apropiadas para el descubrimiento de factores asociados a la deserción estudiantil utilizando herramientas de software libre.

Evaluar los resultados obtenidos en los programas de pregrado de Ciencias Naturales y Ciencias Sociales y Humanas, con el fin de determinar el conocimiento acerca de los factores asociados a la deserción estudiantil en la Universidad de Nariño Sede Pasto.

Preparar el informe final de investigación y la socialización de los resultados obtenidos.

2. MARCO TEÓRICO

2.1. Deserción Estudiantil

Actualmente, la deserción se considera como una problemática compleja y en continua discusión, sin embargo, se ha detectado que es un fenómeno que tiene múltiples causas: individuales, académicos, socioeconómicos e institucionales.

2.1.1. Definición de deserción

Tinto [33], afirma que en muchas ocasiones la definición de deserción depende de la problemática que se quiera abordar, por lo que puede quedar a criterio del investigador según los requerimientos del tópico a tratar y los objetivos que se deban cumplir. Debido a que no es posible cubrir en su totalidad la complejidad del fenómeno, esto requiere la consideración de una gran variedad de perspectivas, así como distintos casos de abandono, puesto que cada uno de estos, hace necesario una atención diferente o puede exigir similitud en el actuar de las instituciones, De ahí, la raíz de la gran dificultad a la que se enfrentan las IES con la deserción.

En [34], la Universidad Pedagógica Nacional define la deserción universitaria como: "El hecho de que un número de estudiantes matriculados no siga la trayectoria normal del programa académico, bien sea por retirarse de ella o por demorar más tiempo del previsto en finalizarla, por repetir cursos o por retiros temporales". Además, muestra que las causas o variables que influyen en la decisión del estudiante de abandonar su carrera profesional en dicha institución, son: económico-laboral, escasa claridad vocacional, y otros. También puede presentarse por cambio de carrera en la misma institución o cambio de institución donde puede continuar con la misma carrera o con otra.

A partir de la definición emanada por el Ministerio de Educación Nacional antes mencionada, se pueden diferenciar dos tipos de abandono en estudiantes universitarios: uno con respecto al tiempo y otro con respecto al espacio [14].

2.1.1.1. Clases de deserción

La deserción con respecto al tiempo se clasifica en precoz, temprana y tardía.

- **Deserción precoz.** Se presenta cuando un individuo siendo admitido a una IES, no realiza su matrícula.

- **Deserción temprana.** Se presenta cuando un individuo abandona los estudios en los primeros semestres de la carrera.

- **Deserción tardía.** Se presenta cuando un individuo abandona los estudios en los últimos semestres de la carrera.

En la Figura 1, se muestran los diferentes tipos de deserción de acuerdo con el momento del recorrido académico en el que se presente.

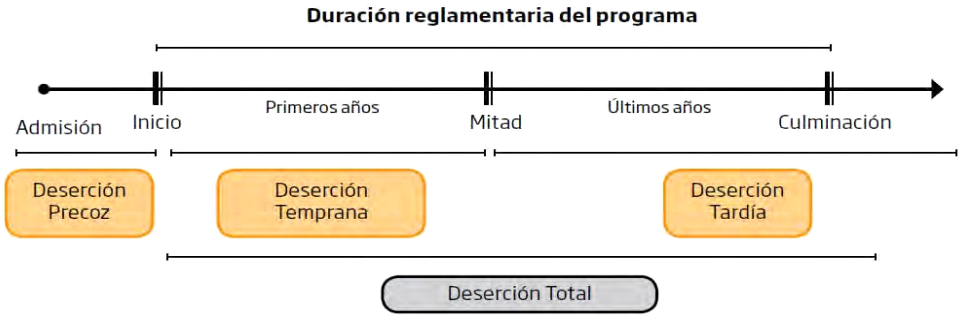

Figura 1. Clasificación de la deserción de acuerdo al tiempo.
Fuente: Deserción estudiantil en la educación superior colombiana. 2009. Pág. 23 [14].

La deserción con respecto al espacio se divide en deserción institucional y deserción interna.

- **Deserción institucional**, la cual, se presenta cuando el estudiante abandona la IES.

- **Deserción interna o del programa académico,** en donde el alumno decide cambiarse a otro programa que ofrece la misma institución [14].

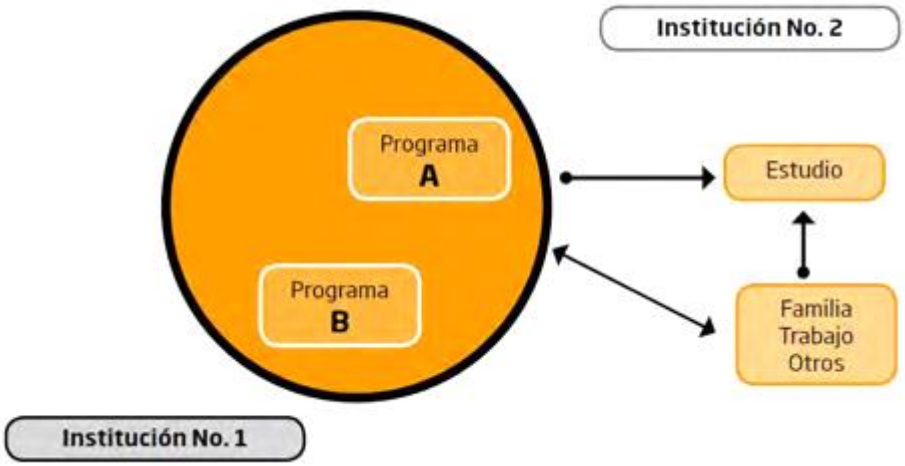

Figura 2. Clasificación de la deserción de acuerdo con el espacio.
Fuente: Deserción estudiantil en la educación superior colombiana. 2009. Pág. 25 [14].

2.1.2. El problema de la deserción

La deserción no es sólo un caso particular de la Universidad de Nariño, más bien, es una situación vigente que afecta a la población estudiantil mundial, siendo un caso a tener en cuenta por entes gubernamentales debido a su impacto en el ámbito económico y social de la nación. De ahí que, es posible destacar investigaciones a nivel internacional, nacional y regional.

2.1.2.1. La deserción a nivel internacional

Se han desarrollado algunos proyectos de investigación aplicando la minería de datos al descubrimiento de patrones de deserción estudiantil. En particular, en la Universidad Nacional de Misiones de Argentina, se realizó un estudio sobre deserción estudiantil utilizando las técnicas de minería de datos. Su objetivo principal fue maximizar la calidad que los modelos tienen para clasificar y agrupar a los estudiantes con base en sus características académicas, factores sociales y demográficos, quienes han desertado de la Carrera Analista en Sistemas de Computación de la Facultad de Ciencias Exactas, Químicas y Naturales, para lo cual analizaron los datos de las cohortes entre los años 2000 y 2006. Encontraron que si bien se realiza una buena clasificación y agrupamiento de las características de los alumnos Activos y Pasivos, salvo el Nivel de Estudio de los Padres, la localidad, el desarraigo y el colegio, no existen otras variables relevantes al análisis socio económico de la deserción estudiantil [19], [20].

En la Universidad Nacional de la Matanza (Argentina) se aplicaron técnicas de minería de datos para evaluar el rendimiento académico y la deserción de los estudiantes del Departamento de Ingeniería e Investigaciones Tecnológicas sobre los datos de los alumnos del periodo 2003 al 2008 [18]. La implementación de este proceso se realizó con el software MS SQL Server para la generación de un almacén de datos, el software SPSS para realizar un pre-procesamiento de los datos y el software Weka (Waikato Environment for Knowledge Analysis) para encontrar un clasificador del rendimiento académico y para detectar los patrones determinantes de la deserción estudiantil [18].

En la Universidad Tecnológica de Izúcar de Matamoros (México) se propuso una investigación para identificar las causas que motivan la deserción de sus estudiantes desde que ingresan. Mediante la técnica minería de datos y la herramienta Weka, encontraron relaciones entre atributos académicos que identifican y predicen la probabilidad de deserción, y propusieron una herramienta para el tutor que le permite predecir la probabilidad de deserción de cualquier alumno en cualquier momento de su estancia escolar. En resumen, muestran que los alumnos de la UTIM desertan por las siguientes tres causas principales: La edad es un factor importantísimo que tiene que ver con la madurez y perspectiva de futuro de los estudiantes, los ingresos familiares, para aquellos alumnos cuya edad sea menor o igual a 18 años, puesto que a esta edad aún dependen de los ingresos familiares para el costo de su educación, y el nivel de inglés, para aquellos alumnos cuya edad sea mayor a 18 años [37].

El deseo de reducir el abandono estudiantil, también ha motivado a investigadores europeos, en particular, Sánchez, J. en 2014, realizó una investigación en la Universidad Politécnica de Madrid, buscando identificar los posibles desertores durante el primer año, y así, implementar diversas políticas en ellos para evitarlo. Para lo cual, tomó datos del Sistema Integrado de Información Universitaria (SIIU) y usando metodología CRISP-DM aplicó algunas técnicas de minería de datos, apoyándose en softwares como Weka, Knime y Clementine, y ejecutó una comparación de los algoritmos SMO, Naive Bayes Tree, Logistic, KNIME Tree, C4.5 y Random Forest, siendo este último el que se destaca del resto [27].

2.1.2.2. La deserción a nivel nacional

Los estudios que se han hecho sobre deserción estudiantil en Colombia han sido planteados por el Ministerio de Educación Nacional, teniendo en cuenta parámetros como la tasa de deserción y el estado socioeconómico del estudiante, pero sus resultados han sido enfocados al planteamiento de estrategias para el financiamiento de créditos por parte del ICETEX, mas no aplicados a la búsqueda de métodos para el mejoramiento del nivel educativo en todas sus etapas o a la determinación de perfiles de estudiantes que estén propensos a abandonar la universidad o caer en bajo rendimiento. Los estudios observados en el país se realizan utilizando encuestas, análisis de datos basado en estadísticas y porcentajes calculados con historiales de datos académicos.

En la Universidad Nacional de Colombia, Sede Medellín, Rico [23] llevó a cabo un estudio con el objetivo de caracterizar la deserción estudiantil de pregrado y suministrar los elementos de análisis necesarios a los responsables de la gestión académica y administrativa sobre las políticas y acciones a emprender en esta materia. Estudió la mayor parte de la población estudiantil que había desertado en el periodo de cinco años comprendido entre los años 2001 y 2005, periodo para el que se disponía de abundante información que facilitó el análisis. Esta investigación reveló que la deserción en la Sede Medellín de la Universidad Nacional es alta, entre el 45% y 50% en las cohortes estudiadas, aunque su tasa promedio acumulada es inferior a la del país que es del 52%. El factor de mayor incidencia en la deserción para dicha institución es la mortalidad académica, cuya participación es en promedio del 60% sobre la deserción total. Esto se debe a que, en su mayoría, los niveles de la formación preuniversitaria de los desertores están por debajo de los promedios presentados por el conjunto de los estudiantes. Las áreas de conocimiento donde es particularmente débil esta formación son las de ciencias exactas y naturales. La deserción por factores no académicos corresponde en promedio al 40% de la deserción general y están relacionados principalmente con causas socioeconómicas. Aunque la Universidad cuenta con un importante número de programas asistenciales y el costo de la matrícula es relativamente bajo, muchos no logran sustentar su permanencia, debido a que la mayoría de la población estudiantil proviene de estratos socioeconómicos bajos.

Castaño et al. [1] llevaron a cabo el estudio “Determinación de la deserción estudiantil en la Universidad de Antioquia”. Esta investigación tuvo como objetivo

analizar la deserción de los estudiantes universitarios, en particular de la Facultad de Ingeniería, establecer los niveles de deserción y sus principales determinantes, con el fin de diseñar políticas que controlen la deserción temprana y tardía. Para ello, se utiliza la metodología de modelos de duración y se evalúa el riesgo de abandono; esta metodología también conocida como modelos de análisis de supervivencia, permite determinar, tanto la probabilidad de que el individuo deserte sujeto al tiempo que ha permanecido vinculado a la universidad, como los principales factores que conllevan a tomar la decisión de abandonar los estudios. Para esta investigación sus autores tomaron los datos del Sistema de Información de Matricula y Registro, del Módulo de Inscripción y Selección Sistemática, y de una encuesta aplicada a los estudiantes de la cohorte 1996-II con el fin de complementar las bases de datos de la universidad, básicamente, en dos categorías: socioeconómica e institucional. El análisis lo realizaron sobre una muestra de 138 estudiantes. Como resultado del estudio se concluyó que el mayor porcentaje de desertores se presenta en los cuatro primeros semestres y en menos proporción después del quinto semestre. En cuanto a la deserción precoz, los porcentajes siguen siendo altos sin que hasta la fecha se haya investigado por sus posibles causas y se hayan tomado las medidas adecuadas para disminuirla.

Malagón et al. [13] desarrollaron la investigación titulada “Estudio de la deserción estudiantil de los programas de pregrado de la Universidad de los Llanos”, con el propósito de determinar el comportamiento de la deserción en esta institución. El estudio consistió en un análisis desde el enfoque institucional a 12 programas de pregrado en cuanto a estudiantes inscritos, matriculados y graduados en las cohortes 1998 a 2004, semestre por semestre. Una vez identificados los desertores se revisaron sus hojas de vida y se hicieron entrevistas vía telefónica. Dentro de las conclusiones del estudio la deserción más alta se presentó en los primeros cinco semestres, con preponderancia en primero, conocida como deserción inicial. El programa con mayor deserción es Ingeniería de Sistemas con 56%. El mayor determinante de la deserción es el bajo rendimiento académico evidenciado en la repitencia de asignaturas, especialmente del área de ciencias básicas.

En la Universidad de La Sabana (Cundinamarca), se realizó un proyecto de investigación donde el objetivo era seleccionar, de una base de datos de estudiantes, los atributos que tuvieran mayor incidencia en la deserción de la Universidad entre los años 2004-2008, con el método de clasificación Rough Sets, utilizando el paquete ROSE2 [21].

Para Salcedo [26], el éxito académico en las universidades está supeditado a una serie de factores internos y externos que afectan notoriamente el rendimiento general de la misma en sus distintos programas. Por lo tanto, las causas que determinan la deserción se pueden atribuir a varios problemas externos e internos a la universidad, problemas intrínsecos al estudiante y a otras causas. Hecho que se evidencia en la Figura 3.

Franco [4], desde una perspectiva psicológica, analizó los factores que intervienen tanto en el ingreso como en la permanencia de los estudiantes de la Universidad de la Sabana, encontrando que estos, están determinados básicamente por la

orientación vocacional recibida en el bachillerato. Además, advierte que la elección de la carrera está influenciada por identificaciones tempranas, por grupos sociales, por valores familiares, por la utilidad percibida de la carrera, así como el prestigio social que la profesión representa. De ahí que, responsabiliza la deserción universitaria a la falta de orientación profesional adecuada para escoger la carrera universitaria a seguir.

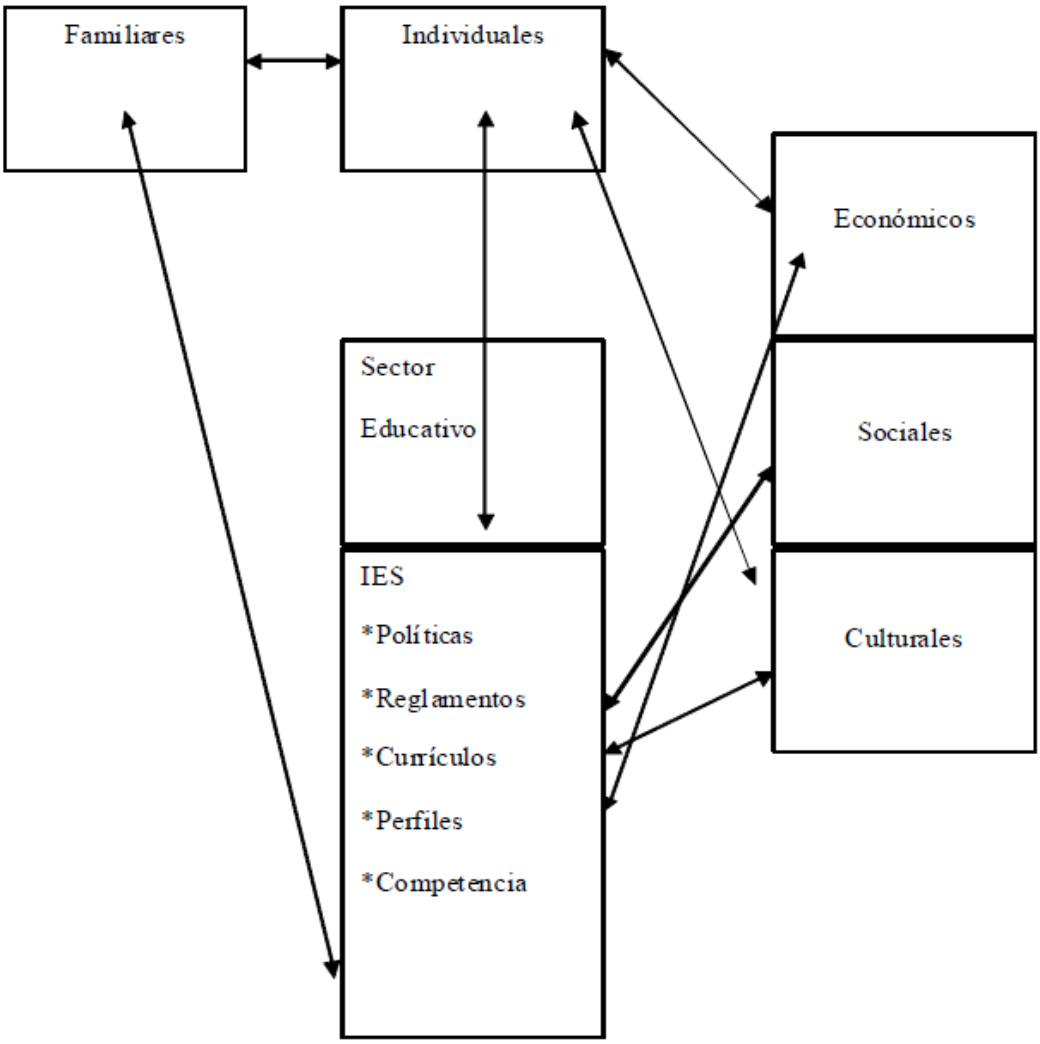

Figura 3. Causas de la deserción.
Fuente: Deserción universitaria en Colombia. 2010. Pág. 57. [26].

El Ministerio de Educación (MEN) apoyado en el Centro de Estudios sobre Desarrollo Económico (CEDE) de la Universidad de los Andes en conjunto con las instituciones de educación superior, el ICFES, el ICETEX, el SISBEN, SNIES y el observatorio laboral para la educación actualizan con frecuencia el sistema SPADIES (Sistema para la Prevención y Análisis de la Deserción en las Instituciones de Educación Superior), el cual, "consolida y ordena información que permite hacer seguimiento a las condiciones académicas y socioeconómicas de los estudiantes que han ingresado a la educación superior en el país. De esta manera, permite conocer el estado y evolución de la caracterización y del rendimiento

académico de los estudiantes, lo cual es útil para establecer los factores determinantes de la deserción, para estimar el riesgo de deserción de cada estudiante y para diseñar y mejorar las acciones de apoyo a los estudiantes orientados a fomentar su permanencia y graduación."[2]

Las variables que SPADIES tienen en cuenta para el análisis de la deserción son:

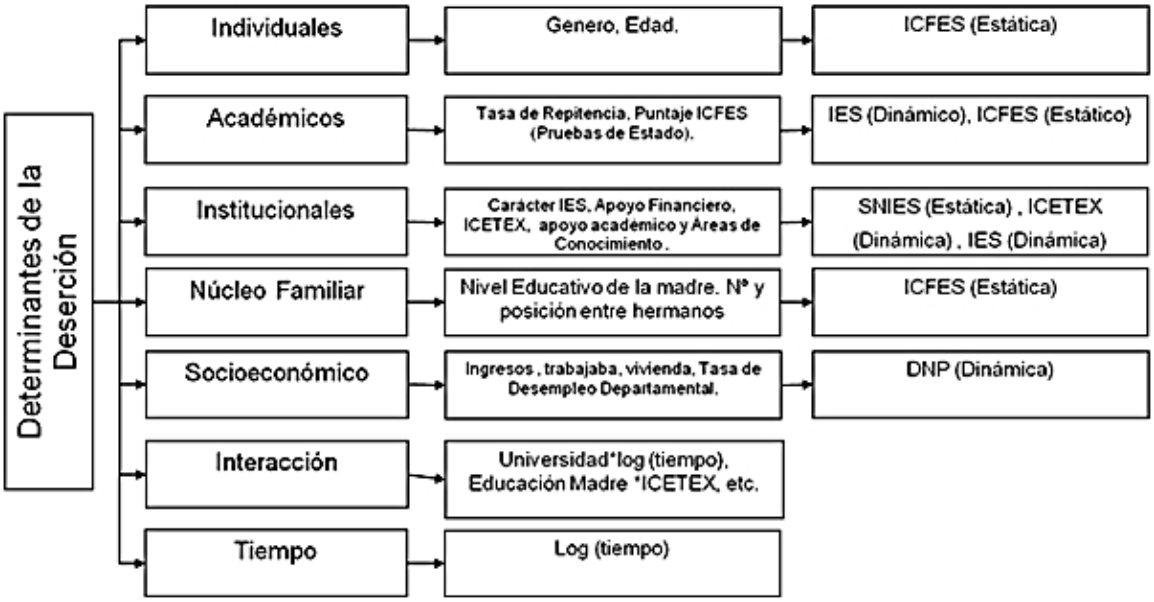

Figura 4. Variables usadas por SPADIES.
Fuente: https://www.mineducacion.gov.co/sistemasdeinformacion/1735/w3-article-254668.html

De acuerdo con el SPADIES, Colombia en los últimos años, ha avanzado en la disminución de la deserción estudiantil en educación superior, puesto que, la meta para la tasa de deserción por periodo (anual) para la vigencia 2015 era 9.7% para el nivel universitario y se logró reducirla a 9.3%. Sin embargo, la tasa de abandono a corte abril de 2016 es del 46,1%, siendo Huila, Caldas y Santander los que menores tasas de deserción presentan. Por su parte, Putumayo, Casanare y La Guajira es donde más registran abandono [28].

Además, para el año 2015 se obtuvo una tasa de graduación del 34,5%, siendo los programas de ciencias de la salud donde mayor acierto se presenta, con un porcentaje del 44,32%, y los programas con menor índice de graduación están el área de Agronomía, veterinaria y afines, con un porcentaje del 24,20%, como se evidencia en la Tabla 1.

[2] Tomado de http://www.mineducacion.gov.co/sistemasdeinformacion/1735/w3-article-254648.html

Área de Conocimiento	Universitaria
Agronomía, veterinaria y afines	10,16%
Bellas artes	8,90%
Ciencias de la educación	9,68%
Ciencias de la salud	5,96%
Ciencias sociales y humanas	8,86%
Economía, administración, contaduría y afines	10,05%
Ingeniería, arquitectura, urbanismo y afines	9,61%
Matemáticas y ciencias naturales	11,06%

Tabla 1. Tasa de Deserción Anual 2015 por Área de Conocimiento y Nivel de Formación. Fuente: SPADIES, Fecha de corte abril de 2016.

2.1.2.3. La deserción a nivel regional

Timarán y Jiménez [32] realizaron una investigación con el fin de detectar perfiles de deserción estudiantil con técnicas de minería de datos en los Programas de pregrado de la Universidad de Nariño y la Institución Universitaria CESMAG. Encontraron como patrón general de deserción estudiantil común para las dos IES el promedio bajo de notas, tener materias reprobadas en los primeros semestres de la carrera y un puntaje promedio bajo en las pruebas de estado. En la Universidad de Nariño, además del patrón general de deserción, incide también el pagar una matrícula promedio alta[3] (mayor que $381504), a pesar que este valor es bajo con relación al valor de las matrículas de otras universidades de la región. La procedencia de la zona sur y de la costa pacífica nariñense constituye un factor asociado a la deserción estudiantil. De igual manera, en la Institución Universitaria CESMAG, además del patrón general de deserción, están los siguientes: pertenecer a la Facultad de Arquitectura y Bellas Artes, tener más de 22 años al momento del ingreso.

De acuerdo con el SPADIES, Nariño presenta una tasa de deserción de 39,9%, la cual es inferior a la tasa nacional [28].

2.1.2.4. La deserción en la Universidad de Nariño

A nivel de la Universidad de Nariño, Timarán [30] dirigió un estudio para identificar perfiles de bajo rendimiento académico y deserción estudiantil aplicando técnicas de minería de datos, para lo cual utilizó la base de datos histórica de los estudiantes de pregrado, compuesta por información personal y académica de 46173

[3] Se considera alta para el tipo de estudiantes que ingresan a la Universidad de Nariño, que son de estratos bajos.

estudiantes entre activos, egresados y retirados, acumulada en un periodo de 18 años. Utilizó la herramienta de minería de datos TariyKDD y aplicó las cinco etapas del proceso de descubrimiento de conocimiento: selección, pre-procesamiento, transformación, minería, interpretación y evaluación de resultados. Este estudio permitió identificar que la mayoría de los estudiantes de primer semestre, provenientes de la zona sur del departamento de Nariño, de estratos socioeconómicos bajos y matriculados en algún programa de la facultad de Ciencias Naturales y Matemáticas o en la facultad de Ciencias Humanas, presentan un bajo rendimiento académico. En las facultades de Ciencias Naturales y Matemáticas y en la de Ciencias Humanas, la mayoría de estudiantes que se retiran no reingresan, mientras que, en la facultad de Ingeniería, la mayoría de estudiantes retirados reingresan.

2.1.3. Metodología CRISP-DM

Según Chapman et al. [2], la Metodología CRISP-DM (Cross- Industry Standard Process for Data Mining) construida a base de experiencias reales de cómo la gente elabora proyectos, pretende proporcionar nuevas ideas a quienes desarrollan estudios de minería de datos. La metodología CRISP-DM está descrita en términos de un modelo de proceso jerárquico, la cual, consiste en un conjunto de tareas descritas en 4 niveles (de lo general a lo específico): Fases, Tareas generales, Tareas especializadas e Instancias de proceso. Ver Figura 5 .

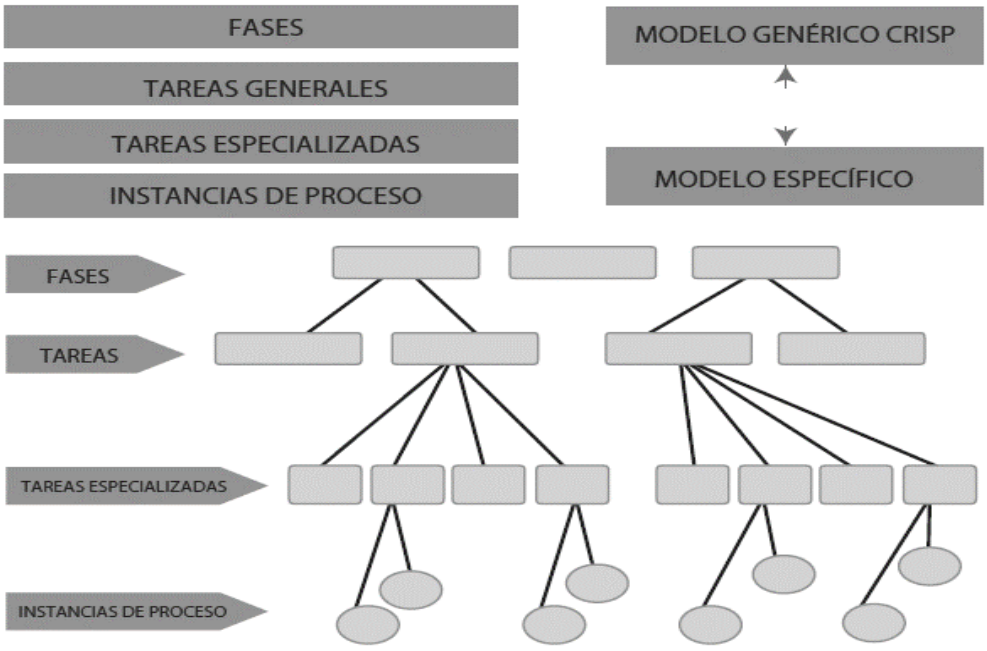

Figura 5. Esquema de los 4 niveles de CRISP-DM.
Fuente: P. Chapman et al.2000. [2].

2.1.3.1. Fases de CRISP-DM

Para Timaran y Jiménez [32], CRISP-DM es la metodología de referencia más utilizada en el desarrollo de proyectos de minería de datos tanto en lo académico como lo industrial. Comprende seis fases[4]: Análisis del problema, análisis de los datos, preparación de los datos, modelado, evaluación y explotación. Ver Figura 6.

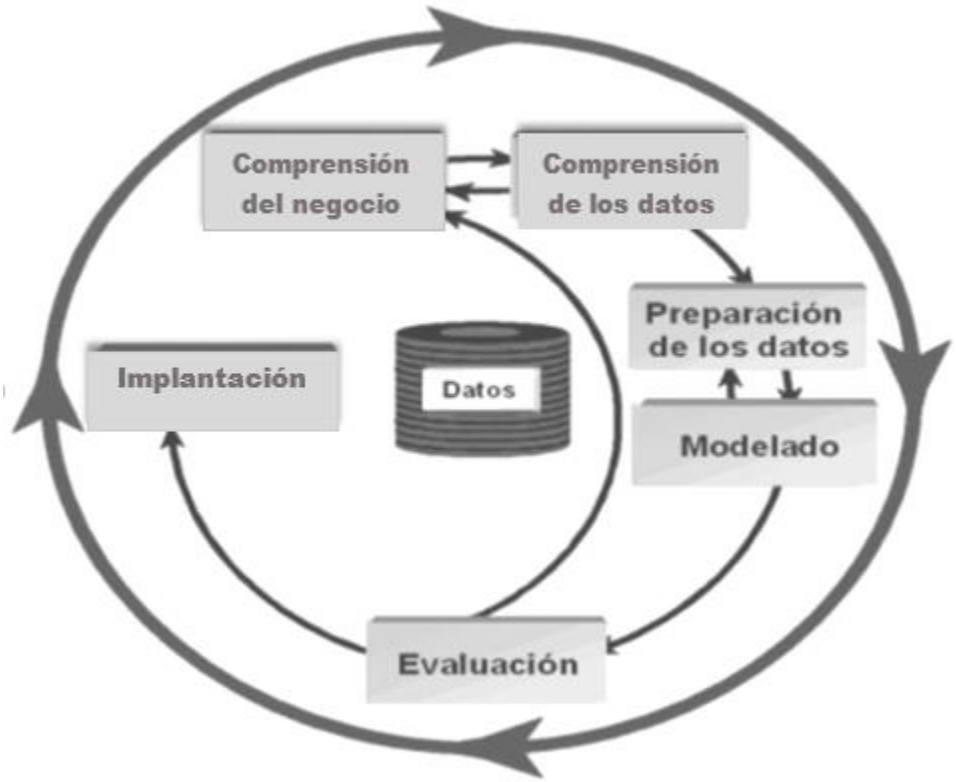

Figura 6. Fases de la metodología CRISP-DM.
Fuente: P. Chapman et al.2000. [2].

2.1.3.1.1. Comprensión del negocio

En esta fase se pretende comprender con exactitud el problema al cual se va a dar solución. Mediante la determinación de objetivos del proyecto, evaluación de la situación actual, establecer los objetivos de la minería de datos y por último generar un plan de proyecto. Ver Figura 7.

[4] Para algunos autores como Chapman et al. [2], las fases se denominan: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación e implantación.

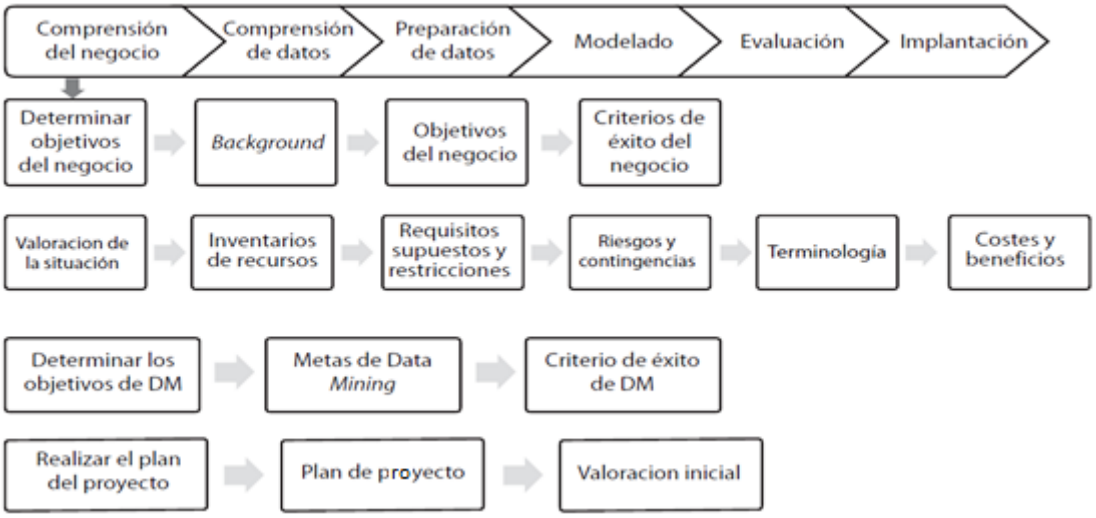

Figura 7. Fases de análisis del problema.
Fuente: P. Chapman et al.2000. [2].

2.1.3.1.2. Comprensión de los datos

El objeto de esta fase es hacer una primera exploración de los datos y reconocer las características más visibles que describen el problema. Dentro de las actividades que se adelantan en esta fase, sobresalen: recolectar datos iniciales, describir los datos, explorar los datos y verificar la calidad de los datos. Ver Figura 8.

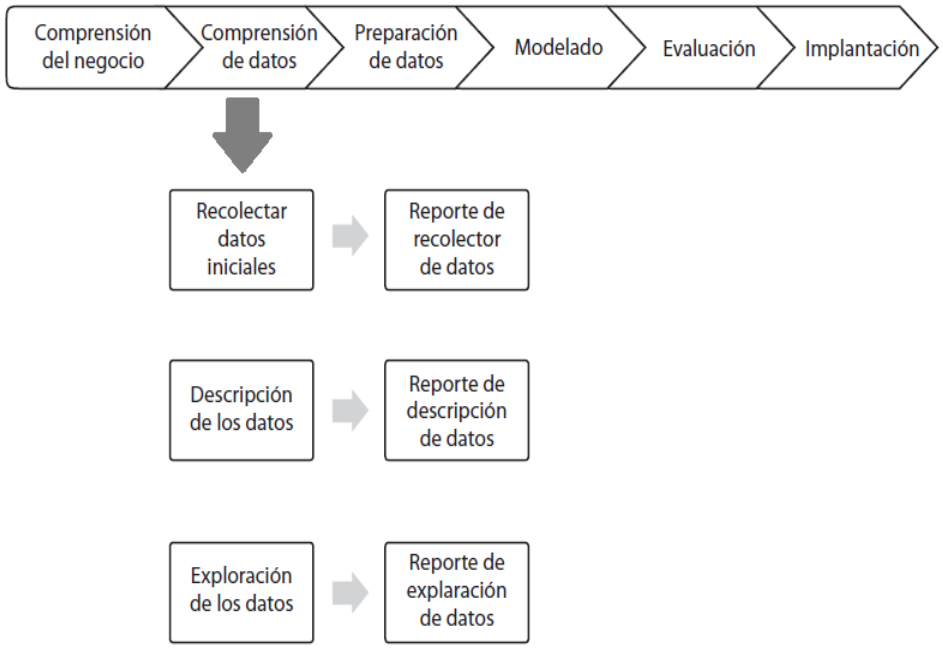

Figura 8. Fase de comprensión de los datos.
Fuente: P. Chapman et al.2000. [2].

2.1.3.1.3. Preparación de los datos

Según Timaran y Jiménez, en [32], “esta fase se usa para adaptarlos a la técnica de minería de datos, mediante la visualización de los datos y la búsqueda de relaciones entre las variables”. Para ello, se realiza procesos de selección, limpieza, estructuración, integración y formateo. Ver Figura 9.

Aquí, se analiza la calidad de los registros, aplicando técnicas estadísticas para la remoción de datos ruidosos, desconocidos, nulos o duplicados. Además, se realiza el tratamiento de registros faltantes o incompletos [5].

El principal objetivo de ésta etapa es detectar y tratar la mayor cantidad de datos inconsistentes, evitando así, extracción inadecuada de conocimiento y la toma decisiones erróneas [5].

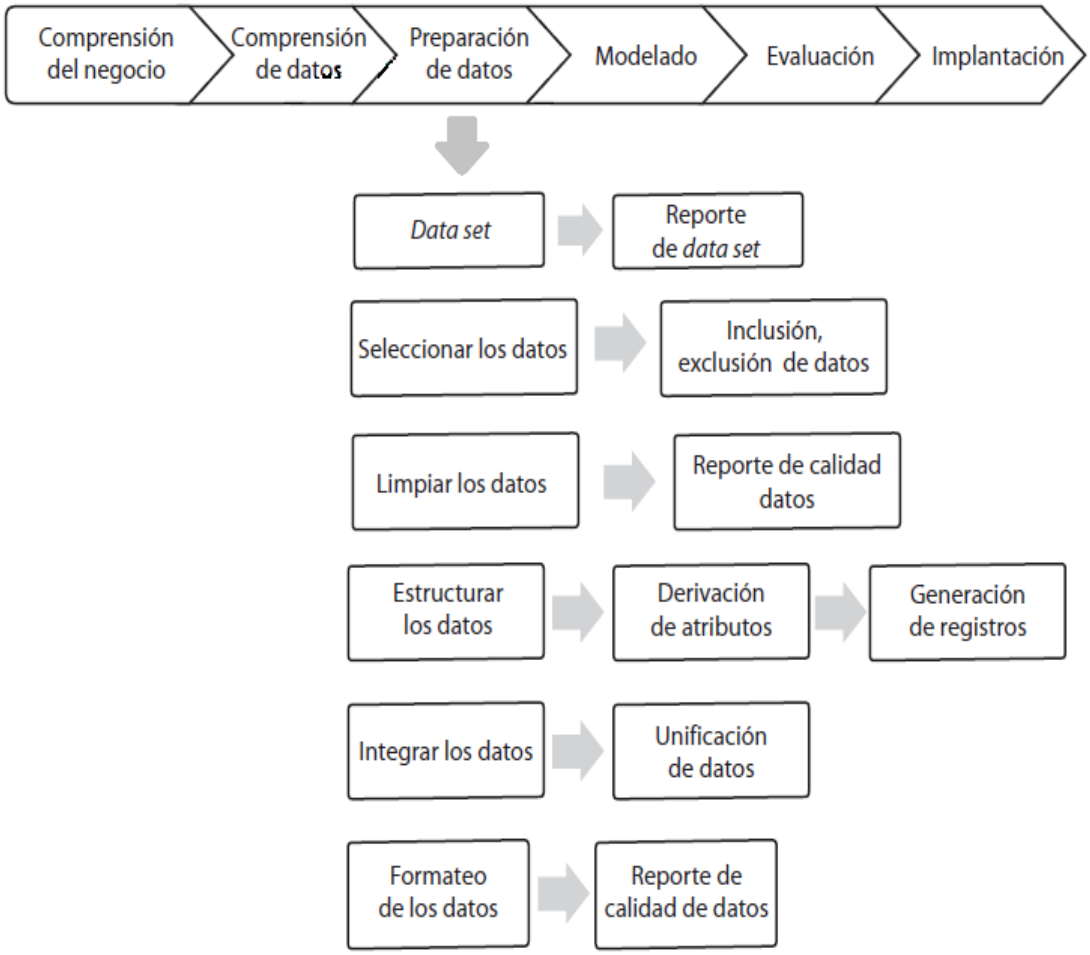

Figura 9. Fase preparación de los datos.
Fuente: P. Chapman et al.2000. [2].

2.1.3.1.4. Modelado

En esta fase se pretende escoger el modelo adecuado de minería, el cual, debe ser apropiado para dar respuesta al problema de investigación. Dentro de las tareas para esta fase, sobresalen: selección de la técnica de modelado, diseño de la evaluación, construcción y evaluación del modelo. Ver Figura 10.

Las técnicas de minería de datos crean modelos que son predictivos o descriptivos. Los modelos predictivos pretenden estimar valores futuros o desconocidos de variables de interés, que se denominan variables objetivo, dependientes o clases, usando otras variables denominadas independientes o predictivas, como por ejemplo pronosticar para nuevos clientes si son buenos o malos basado en su estado civil, edad, género y profesión o determinar para nuevos estudiantes, si desertan o no en función de su zona de procedencia, facultad, estrato, género, edad y promedio de notas. Entre las técnicas predictivas están: clasificación y regresión.

Los modelos descriptivos identifican patrones que explican o resumen los datos y sirven para explorar las propiedades de los aspectos examinados, no para predecir nuevos datos. Por ejemplo, reconocer grupos de personas con gustos similares o identificar patrones de compra de clientes en una determinada zona de la ciudad. Entre las técnicas descriptivas se cuentan: reglas de asociación, patrones secuenciales, clustering y correlaciones.

Por lo tanto, la escogencia de un algoritmo de minería incluye: la selección de los métodos a aplicar en la búsqueda de patrones en los datos, así como la decisión sobre los modelos y los parámetros más apropiados, dependiendo del tipo de aspecto (categóricos, numéricos) a utilizar.

De acuerdo con Han y Kamber [9] y Sattler y Dunemann [29] la clasificación por árboles de decisión es el modelo más utilizado por su simplicidad y facilidad para su entendimiento. El conocimiento obtenido en el proceso de aprendizaje se representa mediante un árbol en el cual cada nodo interior contiene una pregunta sobre un atributo concreto (con un hijo por cada posible respuesta) y cada hoja del árbol se refiere a una decisión (una clasificación). Durante la etapa de construcción del árbol, en forma recursiva, cada conjunto de datos se divide en subconjuntos de acuerdo a un criterio de particionamiento, con el fin de escoger el atributo que mejor separe los ejemplos restantes en clases individuales. Seleccionar el mejor punto de particionamiento es la parte de la construcción del árbol que mayor tiempo consume [29].

Antes de construir un modelo se debe definir un procedimiento para probar la calidad del modelo y su validez. Por tanto, para entrenar y probar un modelo de clasificación, el diseño de prueba específica divide los datos en dos conjuntos: entrenamiento y prueba. Existen diferentes medidas de evaluación del clasificador en la herramienta de minería de datos Weka:

-Usar el conjunto de datos de entrenamiento (Use training set): se emplea todo el conjunto de datos para entrenar el modelo y después se prueba (esta técnica puede ser muy buena para ese conjunto de datos, pero puede ser poco precisa para nuevos datos) [31].

-Proveer un conjunto de datos de prueba (Supplied test set): se emplea un conjunto de datos independiente para entrenar y otro conjunto de datos con los que se está trabajando para prueba (corriendo el riesgo que el conjunto de prueba no refleje o se corresponda con las características de los datos que se emplearon para entrenar el modelo) [31].

-Porcentaje de Partición (Percentage Split): se emplea un porcentaje aleatorio de datos para entrenar y otro porcentaje para probar, este método difiere del anterior en que ambos conjuntos pertenecen al universo de datos con el que se está trabajando por lo que se elimina el riesgo que corre el anterior [31].

-Validación cruzada (Cross validation): Este mecanismo permite reducir la dependencia del resultado del experimento en el modo en el cual se realiza la partición [8]. Para este caso particular se utiliza el método de evaluación validación cruzada con n pliegues (n-fold cross validation). En Weka, Cross validation es la opción por defecto y la más comúnmente utilizada. Este método consiste en dividir el conjunto de entrenamiento en n subconjuntos disjuntos de similar tamaño llamados pliegues (folds) de forma aleatoria. El número de subconjuntos se puede introducir en el campo denominado Folds. Posteriormente se realizan n iteraciones (igual al número de subconjuntos definido), donde en cada una se reserva un subconjunto diferente para el conjunto de prueba y los restantes n-1 (uniendo todos los datos) para construir el modelo (entrenamiento). En cada iteración se realiza un cálculo de error. Por último, se construye el modelo con todos los datos y se obtiene su error promediando los obtenidos anteriormente. Otra ventaja de la validación cruzada es que la varianza de los n errores de muestra parciales, permite estimar la variabilidad del método de aprendizaje con respecto al conjunto de datos. Comúnmente, se suelen utilizar 10 particiones (10-fold cross validation) [10].

Por otra parte, es bastante sencillo evaluar o estimar el coste de un clasificador para un determinado conjunto de ejemplos si se dispone de la matriz de confusión. La matriz de confusión (Confusion Matrix) representa de forma detallada el número de instancias que son predichas por clase. La suma de los registros que se representan en cada fila i, i =1…n constituyen el número de instancias que realmente pertenecen a la clase i. similarmente la sumatoria de los ejemplos o registros en cada columna j, j = 1…n son las instancias que ha predicho el algoritmo al valor j de la clase. Los valores en la diagonal son los aciertos y el resto son los errores de clasificación (ejemplos que pertenecían a la clase i de la fila i y fueron clasificados incorrectamente en otra) [10]

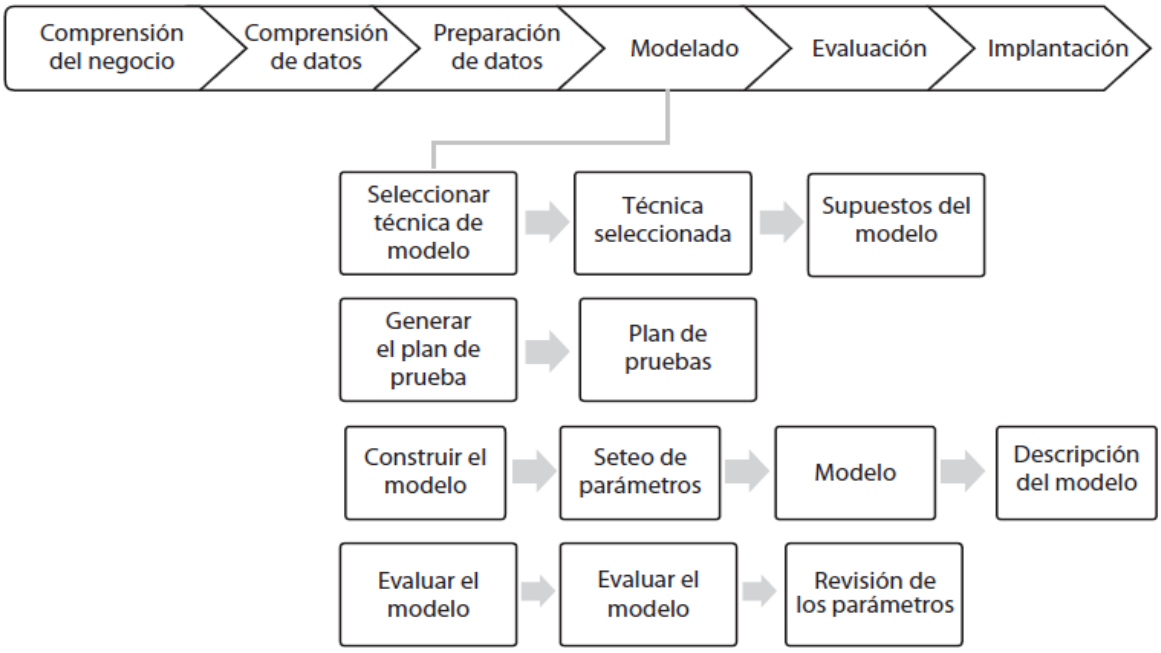

Figura 10. Fase de modelado.
Fuente: P. Chapman et al.2000. [2].

2.1.3.1.5. Evaluación

Se examina el modelo teniendo en cuenta las condiciones de éxito del problema. Buscando siempre su utilidad a las necesidades del negocio o investigación. Según Timarán Pereira, en [32], en esta etapa se descifran los patrones descubiertos y posiblemente se retorna a las anteriores etapas para posteriores iteraciones. Puede incluir la visualización de los patrones extraídos, la remoción de los que son redundantes o irrelevantes y la traducción de los patrones útiles en términos que sean entendibles para el usuario. Por otra parte, se consolida el conocimiento descubierto para incorporarlo en otro sistema para posteriores acciones o, simplemente, para documentarlo y reportarlo a las partes interesadas; también, para verificar y resolver conflictos potenciales con el conocimiento previamente descubierto. Esta fase comprende tareas como: evaluar los resultados, revisión del proceso y determinación de nuevos pasos. Ver Figura 11.

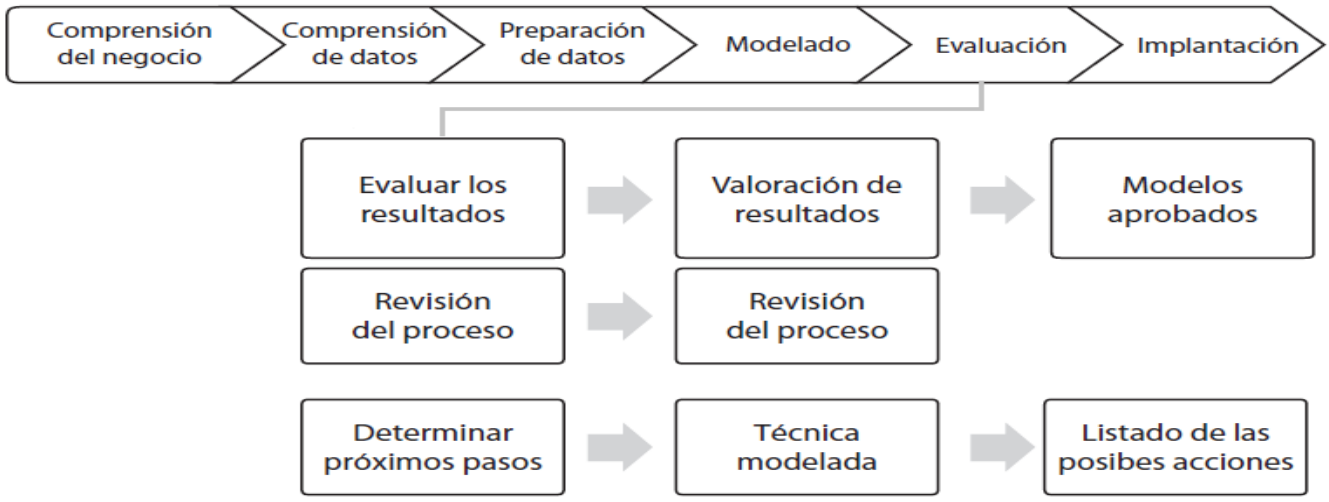

Figura 11. Fase de evaluación.
Fuente: P. Chapman et al.2000. [2].

2.1.3.1.6. Implantación

En esta fase se ve reflejado el trabajo del investigador, pues es aquí, donde el conocimiento obtenido de los datos se transforma en acciones dentro del negocio. Comprende las tareas: plan de implantación, plan de monitoreo, producción del informe final y revisión del proyecto. Ver Figura 12.

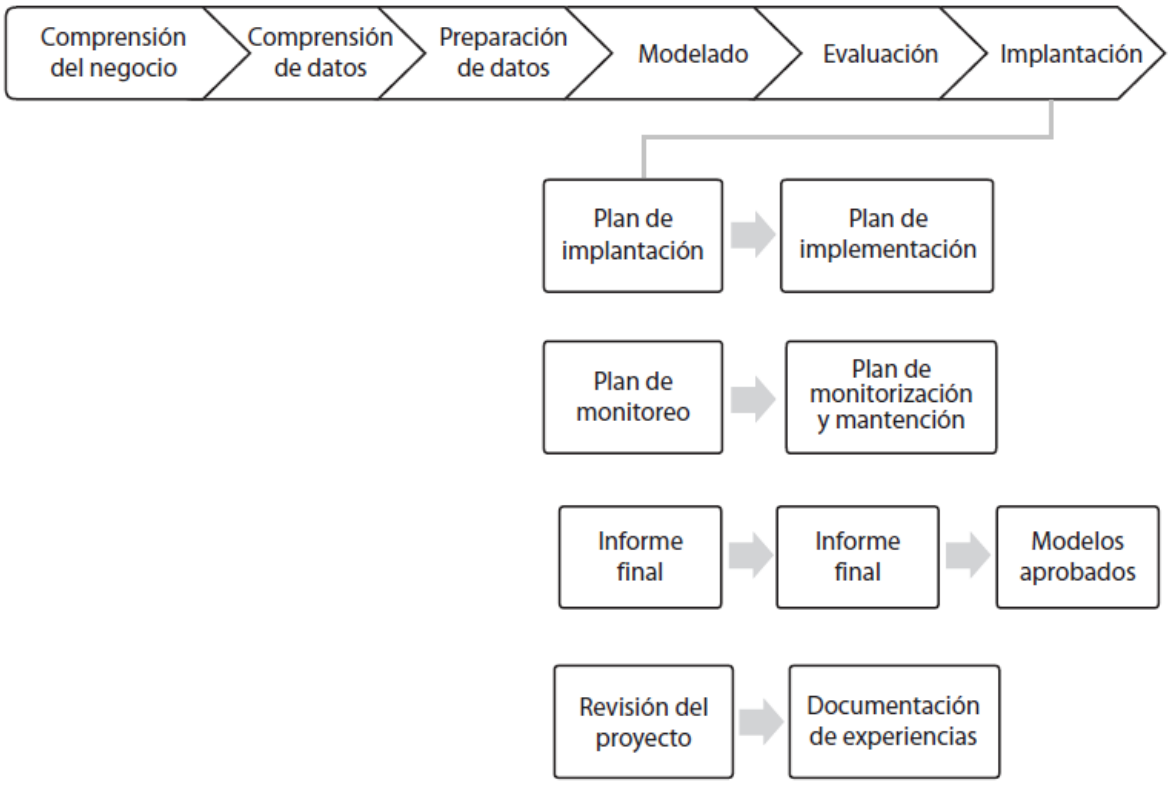

Figura 12. Fase de Implantación.
Fuente: P. Chapman et al.2000. [2].

3. CRISP-DM EN DESERCIÓN ESTUDIANTIL EN LA UNIVERSIDAD DE NARIÑO

La investigación fue de tipo descriptivo bajo el enfoque cuantitativo, aplicando un diseño no experimental.

Para el desarrollo de este trabajo, el punto de referencia son las fases de la metodología CRISP-DM (*Cross Industry Standard Process for Data Mining*), como base para la detección de factores asociados a la deserción estudiantil en los programas profesionales de la Universidad de Nariño Sede Pasto. Se trata de un modelo de proceso que describe los enfoques comunes que utilizan los expertos en minería de datos. CRISP-DM es uno de los modelos utilizados, principalmente, en los ambientes académico e industrial y la guía de referencia más ampliamente usada en el desarrollo de este tipo de proyectos [10].

3.1. Compresión Del Negocio

3.1.1. Objetivo del negocio

Descubrir factores asociados a la deserción estudiantil en los programas de pregrado de la Universidad de Nariño sede Pasto, a partir de la clasificación de cada programa como Ciencias Naturales o Ciencias Sociales y Humanas, y de datos socioeconómicos, académicos, institucionales y de admisión registrados en las bases de datos de la Universidad, utilizando técnicas estadísticas y de minería de datos.

3.1.2. Valoración de la situación actual

Se cuenta con estudios previos en cuanto a deserción universitaria a nivel internacional, nacional y regional. Además, el Centro de Informática de la Universidad de Nariño, posee información de tipo personal, socioeconómico, académico e institucional, de los estudiantes admitidos entre 2006 y 2017. Por tal razón, es posible crear dos repositorios, uno histórico de 10 años y uno por cohortes con una ventana de observación de 6 años para llevar a cabo el proyecto. Por otro lado se cuenta con herramientas de Software libre (SGBD postgreSQL y Weka) para el tratamiento y modelado de datos.

3.1.3. Objetivos de Data Mining

Identificar los factores socioeconómicos, académicos e institucionales asociados a la deserción estudiantil de los programas de pregrado de la Universidad de Nariño discriminados como Ciencias Naturales o como Ciencias Sociales y Humanas.

3.2. Comprensión De Los Datos

En esta fase se identificó y se familiarizó con los datos, permitiendo equiparar los atributos más relevantes para la investigación tales como: periodo de ingreso, fecha de ingreso, nombre de la facultad, nombre de carrera, notas, entre otros.

Por otra parte, mediante la utilización de la herramienta PostgreSQL fue posible constatar que se trata de una información completa y de calidad apta para el procesamiento y limpieza de los datos.

3.2.1. Descripción de los datos

Las tres bases adquiridas presentaron las siguientes características: La base Estudiantes registró 29.317 datos y 78 variables, incluidos los estudiantes de las diferentes extensiones con las que cuenta la Universidad de Nariño, duplicados, nulos, etc. En ella, se encontraron datos personales, socioeconómicos, de admisión y académicos. La base Pagos registró 183.145 datos y 7 variables, donde se muestran los pagos realizados por los distintos estudiantes en cada uno de los periodos que estuvo matriculado. La base Notas registró 980.324 datos y 6 variables, donde se anotaron las calificaciones de cada una de las materias cursadas por los diferentes estudiantes.

Con el fin de comprender cada una de las bases adquiridas se creó un diccionario de datos de cada una de las tablas de la base de datos. Como se muestra en las Tablas del Anexo A.

Puesto que las bases iniciales presentaron datos que no concernían a la presente investigación, se descartaron los registros de estudiantes pertenecientes a extensiones de la Universidad de Nariño distintas a la Sede Pasto, de igual forma aquellos que referían a programas de Posgrado y datos correspondientes a estudiantes ingresados en los años 2016 y 2017.

Dado que el trabajo se orientó a la detección de patrones que determinan la deserción en los programas de pregrado de la Universidad de Nariño sede Pasto clasificados como Ciencia Natural y Ciencia Social se creó la Tabla 2 que contiene la clasificación de los programas de pregrado de la Universidad de Nariño, con el propósito de aclarar que programas pertenecen a las Ciencias Naturales y cuales a las Ciencias Sociales y Humanas para nuestro estudio de caso.

La Universidad de Nariño cuenta con 11 facultades y 37 programas de pregrado clasificados como se ilustra en la siguiente tabla.

CIENCIAS NATURALES	CIENCIAS SOCIALES Y HUMANAS
Facultad de ciencias exactas y naturales Biología. Física. Licenciatura en informática. Licenciatura en matemáticas. Química. **Facultad de ciencias agrícolas** Ingeniería agroforestal. Ingeniería agronómica. Ingeniería ambiental. **Facultad de ingeniería** Ingeniería de sistemas. Ingeniería electrónica. Ingeniería civil. **Facultad de ciencias pecuarias** Ingeniería en producción acuícola. Medicina veterinaria. Zootecnia. **Ingeniería agroindustrial** Ingeniería agroindustrial.	**Facultad de educación** Licenciatura en lengua castellana y literatura. Licenciatura en educación básica con énfasis en ciencias naturales y educación ambiental. **Facultad de ciencias de la salud** Medicina. **Facultad de derecho y ciencias políticas** Derecho. **Facultad de ciencias humanas** Geografía. Licenciatura en ciencias sociales. Licenciatura en educación básica con énfasis en humanidades, lengua castellana e inglés. Licenciatura en filosofía y letras. Licenciatura en inglés y francés. Psicología. Sociología. Licenciatura en educación básica con énfasis en ciencias sociales. (Últimos admitidos en 2015 A) **Facultad de artes** Arquitectura. Diseño gráfico. Licenciatura en música. Diseño industrial. Licenciatura en artes visuales. **Facultad de ciencias económicas y administrativas** Economía. Administración de empresas. Mercadeo. Contaduría pública. Comercio internacional.

Tabla 2. Clasificación programas Universidad de Nariño.
Fuente: Elaboración propia.

3.2.2. Exploración de los datos

Como primera instancia de la investigación se creó el repositorio histórico, con el fin de hacer una exploración de los datos y obtener unas estadísticas descriptivas de los estudiantes que ingresaron desde el periodo académico 2006A hasta 2015B.

El repositorio histórico fue la fuente inicial para la eliminación y agregación de nuevos atributos que permitió crear un repositorio por cohortes apto para la preparación de los datos y aplicación de las técnicas de minería.

Durante el periodo comprendido entre 2006A hasta 2015B, ingresaron 17510 estudiantes, los cuales se distribuyen por facultad como se muestra en la Tabla 45 (Anexo B). Se observa que, la facultad con mayor número de admitidos es Ciencias Humanas y la facultad con menor número de estudiantes ingresados es Ingeniería Agroindustrial, debido a la cantidad de programas académicos con que cuentan.

De los datos, el número de estudiantes pertenecientes a los programas clasificados como Ciencias Naturales son 6802 estudiantes equivalente al 38,84%, y 10708 estudiantes pertenecen a los programas clasificados como Ciencias Sociales y Humanas equivalente al 61,16% (Tabla 3).

TIPO CIENCIA	No DE ESTUDIANTES	%
Ciencias Naturales	6802	38,84
Ciencias Sociales	10708	61,16
TOTAL	17510	100

Tabla 3. Estudiantes clasificados como Ciencias Naturales y Ciencias Sociales.
Fuente: Elaboración propia.

Mediante la Tabla 46 (Anexo B), se pudo evidenciar el número de estudiantes que han ingresado por periodo desde año 2006 hasta el año 2015. Donde se resalta el hecho que en los primeros semestres de los años 2006 hasta 2010 ingresan pocos estudiantes, pero, a partir del periodo 2010B el número de estudiantes admitidos por periodo es similar, debido a que la Universidad de Nariño realiza admisiones a cada programa anualmente, en consecuencia, se vio afectada por el cambio de Calendario B a Calendario A en los colegios públicos del departamento de Nariño, y las directivas deciden modificar el periodo de admisión para ciertos programas.

Ahora, de los estudiantes que ingresaron desde el periodo 2006A hasta el periodo 2015B, mediante la Tabla 4 se observó que el 7,74% son egresados, y el 23,47% son graduados.

Por otra parte, para el periodo comprendido desde 2014A hasta 2015B no existe estudiantes egresados ni graduados como se evidencia en la siguiente tabla, debido a que los programas profesionales de pregrado constan de diez (10) semestres.

PERIODO ACADEMICO	TIPO DE EGRESADO			TOTAL
	No Egresado	Egresado	Graduado	
2006A	119	7	123	249
2006B	695	52	669	1416
2007A	125	8	104	237
2007B	755	59	600	1414
2008A	110	4	105	219
2008B	671	89	598	1358
2009A	98	4	91	193
2009B	730	139	606	1475
2010A	56	3	85	144
2010B	494	84	333	911
2011A	520	124	273	917
2011B	558	147	270	975
2012A	644	172	149	965
2012B	624	243	84	951
2013A	693	221	20	934
2013B	1044	2	0	1046
2014A	988	0	0	988
2014B	1003	0	0	1003
2015A	1156	0	0	1156
2015B	959	0	0	959
TOTAL	12042	1356	4110	17510

Tabla 4. Estudiantes clasificados de acuerdo a la variable Egresado.
Fuente: Elaboración propia.

De las bases de datos proporcionadas por el centro de informática se investigó que programas hasta el año 2015, obtuvieron la acreditación académica de alta calidad, con la intención de observar, si este es o no un factor influyente a la hora de indagar sobre deserción.

No	PROGRAMA ACADÉMICO	RESOLUCIÓN ACREDITACIÓN DE ALTA CALIDAD
1	Zootecnia	2754 de 07 de noviembre de 2001
2	Ingeniería Agronómica	2162 de 21 de noviembre de 2001
3	Ingeniería Agroforestal	6288 de 13 de octubre de 2006
4	Biología	382 de 2 de febrero de 2007
5	Física	3566 de 16 de junio de 2008
6	Licenciatura en Educación Básica énfasis en Ciencias Naturales y Educación Ambiental	3603 de 2 de junio de 2009
7	Licenciatura en Lengua Castellana y Literatura	5612 de 25 de agosto de 2009
8	Ingeniería Agroindustrial	5613 de 25 de agosto de 2009
9	Ingeniería de Sistemas	6797 de 6 de agosto de 2010
10	Psicología	6804 de 6 de agosto de 2010
11	Química	1237 de 21 de febrero de 2011
12	Ingeniería Civil	1236 de 21 de febrero de 2011
13	Ingeniería Agronómica	1956 de 28 de febrero de 2013
14	Economía	4560 de 25 de abril de 2013
15	Ingeniería Agroforestal	7755 de 26 de mayo de 2014
16	Licenciatura en Educación Básica con énfasis en humanidades, Lengua Castellana e Inglés	581 de 09 de enero de 2015
17	Psicología	583 de 09 de enero de 2015
18	Ingeniería Agroindustrial	9233 de 26 de junio de 2015
19	Licenciatura en Educación Básica con énfasis en Ciencias Naturales y Educación Ambiental	13751 de 02 de septiembre de 2015
20	Licenciatura en Matemáticas	13752 de 02 de septiembre de 2015
21	Licenciatura en Lengua Castellana y Literatura	13753 de 02 de septiembre de 2015
22	Zootecnia	14315 de 07 de septiembre de 2015
23	Ingeniería Electrónica	20128 de 10 de diciembre de 20015

Tabla 5. Programas académicos con acreditación de alta calidad.
Fuente: http://acreditacion.udenar.edu.co/programas-acreditados

Mediante el sistema gestor de bases de datos (SGBD) postgreSQL, se analizó los datos contenidos en cada una de las variables que componen el repositorio histórico, con el propósito de determinar los atributos más relevantes y aquellos que permanecerían para la construcción del repositorio por cohortes, para ello, se calculó el porcentaje de datos nulos, como se observa en la Tabla 47 (ANEXO B).

Al indagar el porcentaje de datos nulos en el repositorio histórico, se detectó que existen 3316 registros correspondientes a estudiantes ingresados entre los periodos 2006A a 2007B, de los cuales no se tiene información para la mayor parte de sus variables sociodemográficas en la base Estudiantes, como se muestra en la Tabla 16 (Anexo B).

Considerando la definición de desertor proporcionada por el Ministerio de Educación Nacional[5], se encontró en el repositorio histórico un índice de deserción del 47.01% correspondiente a 8232 estudiantes, los cuales, distribuidos por facultades se presentan en la Tabla 6.

FACULTAD	ESTUDIANTES	DESERTORES	%
ARTES	2779	1320	47.50%
CIENCIAS AGRICOLAS	1266	483	38.15%
CIENCIAS DE LA SALUD	571	232	40.63%
CIENCIAS ECONOMICAS Y ADMINISTRATIVAS	2045	785	38.39%
CIENCIAS EXACTAS Y NATURALES	2277	1582	69.48%
CIENCIAS HUMANAS	3226	1453	45.04%
CIENCIAS PECUARIAS	1414	761	53.82%
DERECHO	1059	250	23.61%
EDUCACION	1028	396	38.52%
INGENIERIA	1364	727	53.30%
INGENIERIA AGROINDUSTRIAL	481	243	50.52%
TOTAL	**17510**	**8232**	**47.01%**

Tabla 6. Desertores por Facultad.
Fuente: Elaboración propia.

Donde se pudo observar, que la Facultad con más desertores es Ciencias Exactas y Naturales y la de menor deserción la Facultad de Derecho.

Agrupando las facultades por el Tipo de Ciencia de acuerdo a la clasificación hecha anteriormente, se encontró la siguiente relación de desertores:

[5] "*Desertor*" es aquel estudiante que, en el momento en que se observa, ha abandonado durante dos o más periodos consecutivos la institución o no registra graduación

TIPO CIENCIA	ESTUDIANTES	DESERTORES	%
NATURALES	6802	3796	55.81%
SOCIALES Y HUMANAS	10708	4436	41.43%
TOTAL	**17510**	**8232**	**47.01%**

Tabla 7. Desertores por Tipo Ciencia.
Fuente: Elaboración propia.

En cuanto al periodo de ingreso, el número de estudiantes desertores se relaciona en la Tabla 8, donde, el porcentaje de mayor deserción corresponde a periodo 2012A y el de menor deserción a 2015B.

PERIODO	ESTUDIANTES	DESERTORES	%
2006A	249	128	51.41%
2006B	1416	773	54.59%
2007A	237	135	56.96%
2007B	1414	799	56.51%
2008A	219	117	53.42%
2008B	1358	728	53.61%
2009A	193	99	51.30%
2009B	1475	828	56.14%
2010A	144	59	40.97%
2010B	911	455	49.95%
2011A	917	525	57.25%
2011B	975	507	52.00%
2012A	965	558	57.82%
2012B	951	404	42.48%
2013A	934	392	41.97%
2013B	1046	389	37.19%
2014A	988	351	35.53%
2014B	1003	304	30.31%
2015A	1156	440	38.06%
2015B	959	241	25.13%
TOTAL	**17510**	**8232**	**47.01%**

Tabla 8. Desertores por periodo de ingreso.
Fuente: Elaboración propia.

3.3. Preparación de los Datos

En esta fase se decidió los datos objetivo para el análisis de Minería.

3.3.1. Selección de datos

Dado que el objetivo de la investigación es descubrir los factores asociados a la deserción en los programas de pregrado clasificados como Ciencia Natural y Ciencia Social, de los 17510 estudiantes se entiende que no todas las cohortes presentan estudiantes graduados, por tanto, fue necesario la toma de cohortes completas con estudiantes graduados, para ello, se adquirió las cohortes comprendidas entre los periodos 2008A hasta 2011B.

No se optó por las cohortes entre 2006A hasta 2010B, como se tenía planeado, ya que los 3316 estudiantes que ingresaron durante los periodos 2006A, 2006B, 2007A y 2007B presentaron en la mayoría de sus variables una gran cantidad de datos nulos y faltantes, que distorsionaban la realidad de la información, como se evidencia en la Tabla 48 (Anexo B).

El repositorio por cohortes se lo denominó ***Cohortes2008A2011B*** y lo componen 6192 registros y 78 atributos, del cual se obtuvo la siguiente información.

FACULTAD	No ESTUDIANTES	%
ARTES	1057	17,07%
CIENCIAS AGRICOLAS	433	6,99%
CIENCIAS DE LA SALUD	223	3,60%
CIENCIAS ECONOMICAS Y ADMINISTRATIVAS	630	10,17%
CIENCIAS EXACTAS Y NATURALES	783	12,65%
CIENCIAS HUMANAS	1117	18,04%
CIENCIAS PECUARIAS	512	8,27%
DERECHO	414	6,69%
EDUCACION	411	6,64%
INGENIERIA	434	7,01%
INGENIERIA AGROINDUSTRIAL	178	2,87%
TOTAL	**6192**	**100%**

Tabla 9. Estudiantes por facultad Cohortes2008A2011B.
Fuente: Elaboración propia.

TIPO CIENCIA	ESTUDIANTES	%
NATURALES	2340	37,8%
SOCIALES Y HUMANAS	3852	62,2%
TOTAL	**6192**	**100,0%**

Tabla 10. Estudiantes clasificados por Tipo Ciencia Cohortes2008A2011B.
Fuente: Elaboración propia.

SEXO	ESTUDIANTES	%
F	2492	40,2%
M	3700	59,8%
TOTAL	**6192**	**100,0%**

Tabla 11. Estudiantes clasificados por sexo Cohortes2008A2011B.
Fuente: Elaboración propia.

PERIODO ACADEMICO	No ESTUDIANTES	%
2008A	219	3,5%
2008B	1358	21,9%
2009A	193	3,1%
2009B	1475	23,8%
2010A	144	2,3%
2010B	911	14,7%
2011A	917	14,8%
2011B	975	15,7%
TOTAL	6192	100,0%

Tabla 12. Estudiantes clasificados por periodo de ingreso Cohortes2008A2011B.
Fuente: Elaboración propia.

Del repositorio seleccionado se descubrió que el 52,28% son estudiantes no egresados, el 9,59% son estudiantes egresados y el 38,10% son estudiantes graduados como se muestra en la Tabla 13.

PERIODO ACADEMICO	NO EGRESADO	%	EGRESADO	%	GRADUADO	%	TOTAL
2008A	110	50,23%	4	1,83%	105	47,90%	219
2008B	671	49,41%	89	6,55%	598	44,00%	1358
2009A	98	50,78%	4	2,07%	91	47,20%	193
2009B	730	49,49%	139	9,42%	606	41,10%	1475
2010A	56	38,89%	3	2,08%	85	59,00%	144
2010B	494	54,23%	84	9,22%	333	36,60%	911
2011A	520	56,71%	124	13,52%	273	29,80%	917
2011B	558	57,23%	147	15,08%	270	27,70%	975
TOTAL	**3237**	**52,28%**	**594**	**9,59%**	**2361**	**38,10%**	**6192**

Tabla 13. Estudiantes de acuerdo a clase de EGRESADO Cohortes2008A2011B.
Fuente: Elaboración propia.

Por otra parte el índice de deserción para las cohortes desde 2008A hasta 2011B fue del 47.00% como se evidencia en la Tabla 14.

PERIODO ACADEMICO	ESTUDIANTES	DESERTOR	%
2008A	219	114	52.05%
2008B	1358	655	48.23%
2009A	193	96	49.74%
2009B	1475	716	48.54%
2010A	144	56	38.89%
2010B	911	405	44.46%
2011A	917	439	47.87%
2011B	975	429	44.00%
TOTAL	**6192**	**2910**	**47.00%**

Tabla 14. Desertores por periodo de ingreso Cohortes2008A2011B.
Fuente: Elaboración propia.

Por otra parte, el índice de deserción por facultad se puede ver en la Tabla 15.

FACULTAD	DESERTORES	ESTUDIANTES	%
ARTES	513	1057	48.53%
CIENCIAS AGRICOLAS	170	433	39.26%
CIENCIAS DE LA SALUD	109	223	48.88%
CIENCIAS ECONOMICAS Y ADMINISTRATIVAS	215	630	34.13%
CIENCIAS EXACTAS Y NATURALES	516	783	65.90%
CIENCIAS HUMANAS	497	1117	44.49%
CIENCIAS PECUARIAS	283	512	55.27%
DERECHO	80	414	19.32%
EDUCACION	174	411	42.34%
INGENIERIA	244	434	56.22%
INGENIERIA AGROINDUSTRIAL	109	178	61.24%
TOTAL	**2910**	**6192**	**47.00%**

Tabla 15. Desertores por facultad Cohortes2008A2011B.
Fuente: Elaboración propia.

De acuerdo con la clasificación: “Tipo Ciencia”, se puede notar en la Tabla 16 que para las cohortes 2008A a 2011B se presentó mayor deserción en los programas pertenecientes a las Ciencias Naturales, con un 56,50%.

TIPO CIENCIA	DESERTORES	ESTUDIANTES	%
NATURALES	1322	2340	56.50%
SOCIALES Y HUMANAS	1588	3852	41.23%
TOTAL	**2910**	**6192**	**47.00%**

Tabla 16. Desertores por Tipo Ciencia Cohortes2008A2011B.
Fuente: Elaboración propia.

Del porcentaje de desertores de las *Cohortes2008A2011B*, se detectó que el 73,09% de los estudiantes desertó en los tres primeros semestres, el 17,04% desertó ente el cuarto y séptimo semestre y el 9,86% entre el octavo y décimo semestre como se muestra en la Figura 13.

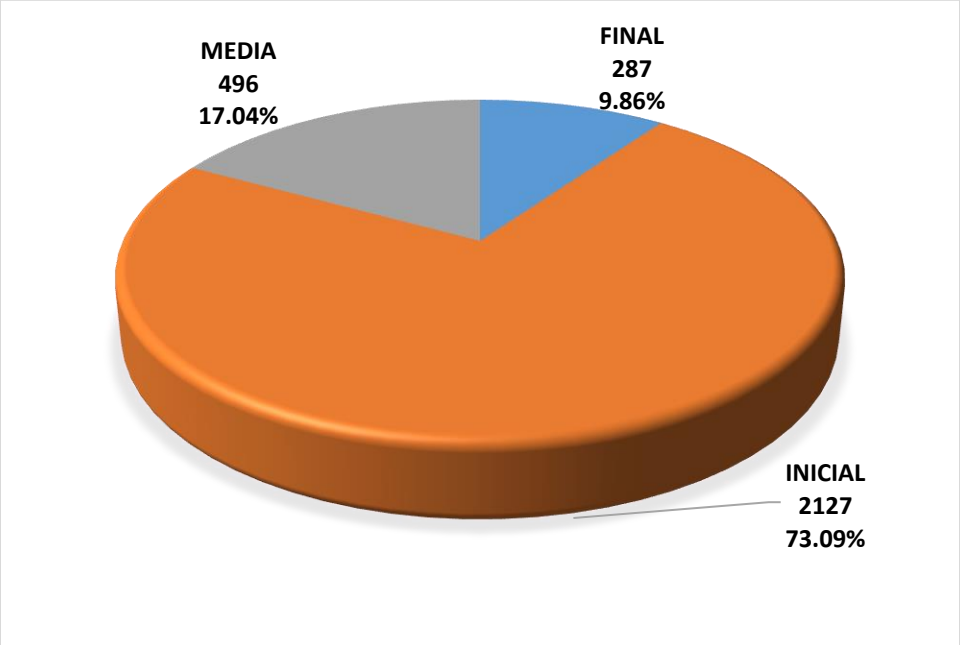

Figura 13. Estudiantes por tipo desertor cohortes 2008A2011B.
Fuente: Elaboración propia.

De los 6192 estudiantes pertenecientes a *Cohortes2008A2011B* el 97,70% de los estudiantes que tuvieron un promedio de notas entre [0-3) son desertores, además, en la medida que aumenta el promedio disminuye la deserción, lo cual, fue un dato relevante para la investigación y de los factores determinantes para el proceso de minería.

PROMEDIO DE NOTAS	DESERTORES	ESTUDIANTES	%
[0-3)	1697	1737	97.70%
[3-3.5)	637	1133	56.22%
[3.5-4)	435	1977	22.00%
[4-5]	141	1345	10.48%
TOTAL	**2910**	**6192**	**47.00%**

Tabla 17. Desertores por promedio de notas Cohortes2008A2011B.
Fuente: Elaboración propia

En cuanto al género se encontró que los hombres son los que más desertan.

SEXO	DESERTORES	ESTUDIANTES	%
F	970	2492	38.92%
M	1940	3700	52.43%
TOTAL	**2910**	**6192**	**47.00%**

Tabla 18. Desertores por Sexo cohortes 2008A2011B.
Fuente: Elaboración propia.

Para la variable programa acreditado, se observó que el 51,07% de los estudiantes que pertenecen a un Programa Acreditado desertan.

PROGRAMA ACREDITADO	DESERTORES	ESTUDIANTES	%
N	1545	3519	43.90%
S	1365	2673	51.07%
TOTAL	**2910**	**6192**	**47.00%**

Tabla 19. Desertores por Programa Acreditado cohortes 2008A2011B.
Fuente: Elaboración propia.

En cuanto a la Zona de procedencia, los que más desertaron fueron aquellos estudiantes de lugares diferentes a Nariño y Putumayo. Por el contrario, los que residen en la Zona Sur del departamento de Nariño son quienes menos registraron abandono, como se observa en la Tabla 20.

ZONA DE PROCEDENCIA	DESERTORES	ESTUDIANTES	%
CENTRO	93	201	46.27%
SUR OCCIDENTE	111	267	41.57%
COSTA	46	90	51.11%
NORTE	167	323	51.70%
OTRA	20	38	52.63%
PASTO	2199	4635	47.44%
PUTUMAYO	102	215	47.44%
SUR	172	423	40.66%
TOTAL	**2910**	**6192**	**47.00%**

Tabla 20. Desertores por Zona de procedencia cohortes 2008A2011B.
Fuente: Elaboración propia.

La Tabla 58 (Anexo D) muestra los municipios que pertenecen a cada una de las zonas de procedencia.

Por otro lado, observando el puntaje de ingreso, se nota que el porcentaje de desertores es similares a cada categoría.

PUNTAJE INGRESOS	DESERTORES	ESTUDIANTES	%
[0,55)	1036	2085	49.69%
[55,61)	942	2162	43.57%
[61,100]	932	1945	47.92%
TOTAL	**2910**	**6192**	**47.00%**

Tabla 21. Desertores por Puntaje de ingreso. cohortes 2008A2011B
Fuente: Elaboración propia.

Para la variable “Tipo de colegio”, ser público o privado tiene porcentajes de deserción similares.

COLEGIO	DESERTORES	ESTUDIANTES	%
DESCONOCIDO	355	688	51.60%
PRIVADO	592	1252	47.28%
PUBLICO	1963	4252	46.17%
TOTAL	**2910**	**6192**	**47.00%**

Tabla 22. Desertores por Tipo de colegio cohortes 2008A2011B.
Fuente: Elaboración propia.

De los estudiantes provenientes del sector rural (estrato cero) desertaron el 49,07%.

ESTRATO	DESERTORES	ESTUDIANTES	%
ALTO	363	847	42.86%
BAJO	1825	3830	47.65%
MEDIO	406	871	46.61%
RURAL	316	644	49.07%
TOTAL	**2910**	**6192**	**47.00%**

Tabla 23. Desertores por Estrato cohorte 2008A2011B.
Fuente: Elaboración propia

En cuanto al valor de la matrícula, los estudiantes que menos desertaron fueron aquellos cuyo valor cancelado fue menor a 6 salarios mínimos legales vigentes diarios.

VALOR MATRICULA	DESERTORES	ESTUDIANTES	%
[0-6)	211	1133	18.62%
[6-8)	1015	1883	53.90%
[8-12)	925	1736	53.28%
12 O MÁS	759	1440	52.71%
TOTAL	**2910**	**6192**	**47.00%**

Tabla 24. Desertores por Valor de matrícula cohortes 2008A2011B.
Fuente: Elaboración propia.

De los estudiantes que perdieron entre 5 y 9 materias desertaron el 67,15%.

MATERIAS PERDIDAS	DESERTORES	ESTUDIANTES	%
0	210	960	21.88%
[1-3)	350	1241	28.20%
[3-5)	661	1160	56.98%
[5-9)	1110	1653	67.15%
9 O MÁS	579	1178	49.15%
TOTAL	**2910**	**6192**	**47.00%**

Tabla 25. Desertores por Materias perdidas cohortes 2008A2011B.
Fuente: Elaboración propia.

Los estudiantes desertores de las ***Cohortes2008A2011B***, registraron la mayor tasa de pérdida en las materias de Formación Básica (53,94%), y la menor tasa en Formación humanística. Aunque, es bueno anotar que tan solo el 30,74% de las materias cursadas por los estudiantes desertores tuvieron un valor menor a 3.0. Ver Tabla 26.

TIPO MATERIA	MATERIAS PERDIDAS	TOTAL MATERIAS	%
BÁSICA	4123	7643	53.94%
COMPLEMENTARIA	221	1086	20.35%
HUMANÍSTICA	2302	11587	19.87%
PROFESIONALIZACIÓN	11046	37243	29.66%
TOTAL	**17692**	**57559**	**30.74%**

Tabla 26. Relación entre desertores y materias perdidas cohortes 2008A2011B.
Fuente: Elaboración propia.

Entiéndase por materias de Formación Básica aquellas referentes a conocimientos esenciales como matemáticas, biología, física, química, etc. Complementarias las que refieren a Electivas, Trabajos de grado, Prácticas, etc. Formación humanística las que involucran deporte formativo, cultura, valores, idioma extranjero, etc. Y como materias de Profesionalización las que son propias de cada programa.

3.3.2. Limpieza de datos

Mediante la utilización SGBD postgreSQL fue posible hacer consultas y análisis de los datos almacenados en *Cohortes2008A2011B*, permitiendo mejorar la calidad de los datos, tales como la detección de nulos, datos fuera de los rangos esperados, datos incoherentes, entre otros. A este tipo de registros se les hizo un estudio minucioso para determinar la forma de corregirlos, reemplazarlos o eliminarlos.

Con el propósito de crear el repositorio por cohortes apto para el proceso de minería se complementó la base de datos adquirida mediante el uso de fuentes de datos externas de acceso público como lo son:

- **SISBEN:** (Sistema de Selección de Beneficiarios Para Programas Sociales). Utilizado para verificar su registro a esta entidad.
- **Registraduría Nacional del Estado Civil.** Utilizada para corrección de números de identificación y fechas de nacimiento.
- **DANE: (**Departamento Administrativo Nacional de estadística). Utilizada para la verificación de códigos que identifican los distintos municipios del país.
- **ICFES:** (Instituto Colombiano para el Fomento de la Educación Superior). Utilizada para la identificación de las materias evaluadas en los distintos años de presentación, como también para la corrección de datos fuera de rango.

En la Tabla 59 (Anexo E) se muestra el proceso de limpieza que se usó para cada una de las variables, como lo es adición, corrección y actualización de datos, así como también, renombrar categorías.

Mediante la utilización del SGBD PostgreSQL se obtuvo el porcentaje de nulos para este nuevo repositorio, con la intención de prestar atención a los atributos con información completa e incompleta para luego proceder a la eliminación y construcción de nuevos atributos.

3.3.3. Construcción de datos

Con la intención de que el repositorio sea apto y completo para las tareas de minería se escogieron los atributos más relevantes que permitieron optimizar el proceso de la investigación, en consecuencia, se crearon atributos a partir de los ya existentes.

La Tabla 44 (Anexo A) describe los atributos agregados que son de vital importancia para el proceso de detección de factores que definen la deserción en los programas de pregrado de la Universidad de Nariño sede Pasto.

Agregados los nuevos atributos en *Cohortes2008A2011B*, se procedió a eliminar las variables que no aportan al proceso de minería, como lo son las de uso temporal, las identificadoras, las correlacionadas o que después del proceso de limpieza fueron reemplazadas por otros atributos, las que presentan altos porcentajes de nulos o de No Aplica (NA), presentan información incoherente, etc. Los campos que se eliminaron y reemplazaron se muestran en la Tabla 60 (Anexo E).

3.3.4. Integración de datos

A partir de las bases de datos Pagos y Notas fue posible la construcción nuevos atributos tales como, semestre actual, semestres cursados, promedio, reingreso, desertor, tipo desertor, smldv (salario mínimo legal diario vigente), homologado, entre otros.

Por otra parte, del repositorio final *Cohortes2008A2011B* se clasificó a los atributos de acuerdo a las categorías descritas en los objetivos del proyecto como son datos socioeconómicos, académicos, disciplinares e institucionales. Sin embargo, se consideró los aspectos disciplinares e institucionales en el mismo grupo de los académicos debido a que la poca cantidad de variables presentes en el aspecto disciplinar e institucional no generaba buenos árboles de clasificación.

Las tablas del Anexo C muestran la clasificación y el número de categorías por variable tanto de los aspectos académicos, disciplinares e institucionales como de los socioeconómicos.

3.3.5. Formateo de datos

En esta etapa se hizo la trasformación de los datos para facilitar el trabajo de minería, se discretizaron algunos atributos de acuerdo al criterio de los investigadores, estas categorizaciones se pueden observar en las tablas que se relacionan en el Anexo C.

Finalmente, los atributos que se consideraron para la determinación de los factores asociados a la deserción en la Universidad de Nariño sede Pasto fueron:

No	ATRIBUTO	CATEGORIAS
1	sexo	F,M
2	vive_familia	S,N
3	actualmente_trabaja	S,N
4	nombre_facultad_la	CIENCIAS HUMANAS, CIENCIAS PECUARIAS, CIENCIAS EXACTAS, CIENCIAS AGRICOLAS CIENCIAS ECONOMICAS Y ADMINISTRATIVAS, ARTES, EDUCACION, INGENIERIA, INGENIERIA INDUSTRIAL, CIENCIAS DE LA SALUD, DERECHO
5	reingreso	S,N
6	desertor	S,N
7	tipo_ciencia	CIENCIAS NATURALES, CIENCIAS SOCIALES Y HUMANAS
8	programa_acreditado	S,N
9	zona_nacimiento	CENTRO, SUR OCCIDENTE, COSTA, NORTE, OTRA, PASTO, PUTUMAYO, SUR, CONTRIBUTIVO SUBSIDIADO
10	regimen_salud	CONTRIBUTIVO, SUBSIDIADO
11	puntaje_ingreso	[0-55), [55-61), [61-100)

12	grupo_familiar	[1-3), [4-5], 6 O MAS
13	ingreso_familiar	[0-5), [5-10), [10-15), 15 O MAS
14	promedio_notas	[0-3], [3-3.5), [3,5-4), [4-5)
15	estado_civil	SOLTERO, NO INFORMA, VIUDO, CASADO, UNION LIBRE
16	colegio	DESCONOCIDO, PRIVADO, PUBLICO
17	jefe_hogar	EL ADMITIDO, MADRE, PADRE
18	edad_ingreso	[0-18], [18-20), 20 O MAS
19	valor_matricula	[0-6), [6-8), [8-12), 12 O MAS
20	materias_perdidas	[1-2], [3-4], [5-8] 9 0 MAS
21	estrato	RURAL, BAJO, MEDIO, ALTO
22	sisben	SIN SISBEN, CON SISBEN

Tabla 27. Atributos usados para los árboles de decisión.
Fuente: Elaboración propia.

Para construir los diferentes árboles de decisión se crearon diferentes repositorios, los cuales se muestran a continuación.

REPOSITORIO	DESCRIPCION	NO.REGISTROS	NO.ATRIBUTOS
General R6192A22	Repositorio final general	6192	22
CNaturales R2340A22	Repositorio general para Ciencias Naturales	2340	22
CSociales R3852A22	Repositorio general para Ciencias Humanas	3852	22

Tabla 28. Repositorios generales
Fuente: Elaboración propia

3.4. Modelado

La minería de datos es la etapa que intenta descubrir patrones de interés hasta el momento desconocidos en un gran volumen de datos, para ello utiliza métodos de clasificación, clustering, patrones secuenciales, reglas de asociación, inteligencia artificial, entre otras.

Para esta fase de modelado se utilizó la herramienta WEKA (*Waikato Environment for Knowledge Analysis*) dado que es un software de libre distribución con un entorno para la experimentación de análisis de datos tales como: la clasificación, asociación, agrupamiento, entre otras. Al ser un software especializado brinda apoyo suficiente para interpretar resultados de manera matemática y estadística y por medio de visualización gráfica o árboles de decisión hace que el uso de esta herramienta de un valor agregado a los resultados obtenidos. Para el desarrollo del proyecto se ensayó la clasificación basado en Bayes con implementación del algoritmo BayesNet y árboles de decisión con la implementación de los algoritmos C4.5, Random Forest y LMT.

3.4.1. Conceptos de los métodos de clasificación

3.4.1.1. Bayes Net

En [3] las redes Bayesianas es una técnica de clasificación y consiste en un modelo gráfico que utiliza arcos para formar una gráfica acíclica y es aplicada en aquellas situaciones en que la incertidumbre se asocia con un resultado que se puede expresar en términos de probabilidad, esta técnica busca determinar relaciones casuales que expliquen un fenómeno y es aplicado en aquellos casos que son de carácter predictivo, por tanto el razonamiento probabilístico o propagación de probabilidades consiste en difundir los efectos de la evidencia por medio de la red para conocer la probabilidad a posteriori de las variables, es decir, a determinadas variables (conocidas) se les otorga una probabilidad y en base a esto se obtiene una probabilidad posterior.

3.4.1.2. C4.5 o J48

En [21] se menciona que este método fue creado por Ross J. Quinlan en 1984 y que también es conocido como J48 en la herramienta Weka. Es un algoritmo de inducción que genera una estructura de reglas o árbol a partir de subconjuntos (ventanas) de casos extraídos del conjunto total de datos de “entrenamiento”. En este sentido, su forma de procesar los datos es parecido al de Id3. El algoritmo genera una estructura de reglas y evalúa su “bondad” usando criterios que miden la precisión en la clasificación de los casos. Emplea dos criterios principales para dirigir el proceso dados por:

1. Calcula el valor de la información proporcionada por una regla candidata (o rama del árbol), con una rutina que se llama “info”.
2. Calcula la mejora global que proporciona una regla/rama usando una rutina que se llama *gain* (beneficio).

Con estos dos criterios se puede calcular una especie de calor de coste/beneficio en cada ciclo del proceso, que le sirve para decidir si crear, por ejemplo, dos nuevas reglas, o si es mejor agrupar los casos de una sola.

El algoritmo realiza el proceso de los datos en sucesivos ciclos. En cada ciclo se incrementa el tamaño de la “ventana” de proceso en un porcentaje determinado respecto al conjunto total. El objetivo es tener reglas a partir de la ventana que clasifiquen correctamente a un número cada vez mayor de casos en el conjunto total.

Cada ciclo de proceso emplea como punto de partida los resultados conseguidos por el ciclo anterior. En cada ciclo de proceso se ejecuta un submodelo contra los casos restantes que no están incluidos en la ventana. De esta forma se calcula la precisión del modelo respecto a la totalidad de datos. Es importante notar que la variable de salida debe ser categórica.

3.4.1.3. Random Forest

En [21] se menciona que Random Forest se basa en el desarrollo de muchos árboles de clasificación. Para clasificar un objeto desde un vector de entrada, se pone dicho vector bajo cada uno de los árboles del bosque. Cada árbol genera una clasificación, el bosque escoge la clasificación teniendo en cuenta el árbol más votado sobre todos los del bosque.

Cada árbol se desarrolla como sigue:

· Si el número de casos en el conjunto de entrenamiento es *N,* prueba *N* casos aleatoriamente, pero con sustitución, de los datos originales. Este será el conjunto de entrenamiento para el desarrollo del árbol.

· Si hay *M* variables de entrada, un número $m<<M$ es especificado para cada nodo, *m* variables son seleccionadas aleatoriamente del conjunto *M* y la mejor participación de este *m* es usada para dividir el nodo. Los valores de m se mantienen constante durante el crecimiento del bosque.

· Cada árbol crece de la forma más extensa posible, sin ningún tipo de poda.

3.4.1.4. *LMT (Logistic Model Tree)*

En [21] LMT proporciona una descripción muy buena de los datos. Un LMT consiste básicamente en una estructura de un árbol de decisión con funciones de regresión logística en las hojas. Como en los árboles de decisión ordinarios, una prueba sobre uno de los atributos es asociado con cada nodo interno. Para enumerar los atributos con k valores, el nodo tiene k nodos hijos, y los casos son clasificados en las k ramas dependiendo del valor del atributo.

Para atributos numéricos, el nodo tiene dos nodos hijos y la prueba consiste en comparar el valor del atributo con un umbral: un caso puede ser clasificar los datos menores en la rama izquierda mientras que los valores mayores en la rama derecha. Un LMT consiste en una estructura de árbol que está compuesta por un juego N de nodos internos o no terminales y un juego de T hojas o nodos terminales.
.

3.4.2. Resultados de clasificación con J48, Bayes Net, Random Forest y LMT

En esta sección los resultados de los algoritmos de clasificación J48, Bayes Net, Random Forest y LMT son presentados y comparados para evaluar cuál es el clasificador más adecuado en la predicción de casos de deserción para la Universidad de Nariño sede Pasto.
Los resultados de los algoritmos mencionados, se obtuvieron a partir de la herramienta WEKA con Cross Validation.
Consecuentemente los resultados de clasificación J48, Bayes Net, Random Forest y LMT son mostrados en la Tabla 29, 30, 31

MODELO	INSTANCIAS CORRECTAMENTE CLASIFICADAS	INSTANCIAS INCORRECTAMENTE CLASIFICADAS	PRECISIÓN	SENSIBILIDAD	TASA DE FN
J48 (C=0,25 M=30)	85,48%	14,51%	0.9033	0.7738	0,2261
Bayesnet	79.40%	20.59%	0.7853	0,7758	0,2268
Random forest	84,56%	15,43%	0.8886	0,7676	0,2323
LMT	85,43%	14,56%	0,9054	0,7704	0,2295

Tabla 29 Resultados para repositorio general R6192A22
fuente: Elaboración propia

MODELO	INSTANCIAS CORRECTAMENTE CLASIFICADAS	INSTANCIAS INCORRECTAMENTE CLASIFICADAS	PRECISIÓN	SENSIBILIDAD	TASA DE FN
J48 (C=0,25 M=30)	84,49%	15,51%	0.9202	0.7942	0,2055
Bayesnet	78.24%	21,75%	0.8203	0,7874	0,2628
Random forest	83,16%	16,83%	0.8784	0,8146	0,1853
LMT	84,78%	15,21%	0,9170	0,8033	0,1966

Tabla 30. Resultados para repositorio CNaturales R2340A22
Fuente: Elaboración propia

MODELO	INSTANCIAS CORRECTAMENTE CLASIFICADAS	INSTANCIAS INCORRECTAMENTE CLASIFICADAS	PRECISIÓN	SENSIBILIDAD	TASA DE FN
J48 (C=0,25 M=30)	85,92%	14,07%	0.8926	0.7487	0,2512
Bayesnet	79.33%	20,66%	0.7528	0,7424	0,2575
Random forest	84,78%	15,21%	0.8801	0,7304	0,2695
LMT	85,82%	14,17%	0,8964	0,7418	0,2581

Tabla 31 resultados para repositorio CSociales y Humanas R3852A22
Fuente: Elaboración Propia

Al hacer la clasificación tanto al Repositorio General como al de Ciencias Naturales y Ciencias Sociales se observa que el modelo j48 es quien mejor clasifica a los estudiantes de cada uno de los repositorios, además al comparar la tasa de falsos negativos (FN), j48 es quien con menos porcentaje los representa, pues en nuestro caso interesa FN con la menor tasa, dado que si quiere proponer políticas de retención estudiantil FN serían los estudiantes que sabiendo que son desertores se los ha considerado como no desertores, por tanto este porcentaje de estudiantes quedarían por fuera de los programas de retención tales como tutorías, u otras de carácter académico, por otra parte si se crea políticas de retención tales como

subsidios de vivienda, trasporte, becas alimenticias u otras debemos tomar el modelo con mayor precisión pues sería el modelo con menor falsos positivos, pues se estaría haciendo una inversión no tan desviada dado que se invertiría inútilmente en tan solo unos pocos estudiantes que sabiendo que no van a desertar se los ha considerado desertores.

En consecuencia, para la determinación de las reglas que pronostican la deserción estudiantil en la Universidad de Nariño Sede Pasto se tomó a partir de los árboles de decisión que genera el modelo J48.

Se crearon árboles de decisión por cada repositorio, donde para la construcción de cada uno de estos fue necesario aplicar criterios de poda de tal manera que no se afectara el criterio de bien clasificados. Las sentencias modificadas en el algoritmo J48 fueron el factor de confianza C (confidence level), y el mínimo número de registros que debían ingresar en cada nodo M. El rango de variación para C se tomó entre el 10% y el 60%, y para M desde 10 hasta 61 registros, a partir de estos criterios de poda, se efectuaron varias iteraciones hasta lograr el máximo porcentaje de bien clasificados, para luego examinar aquellas ramas que sobrepasen un nivel de confianza del 70%, y soporte mínimo del 0,01%.

3.4.3. Arboles de decisión con J48

Para la construcción de los árboles se tomó como clase el atributo DESERTOR el cual, determina si el estudiante es desertor o no los arboles generados son los siguientes.

3.4.3.1. Árbol de decisión para repositorio general R6192A22

```
Time taken to build model: 0.22 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        5293               85.4813 %
Incorrectly Classified Instances       899               14.5187 %
Kappa statistic                          0.7062
Mean absolute error                      0.2171
Root mean squared error                  0.3316
Relative absolute error                 43.5864 %
Root relative squared error             66.4324 %
Total Number of Instances             6192

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,774    0,073    0,903      0,774    0,834      0,713    0,897     0,905     S
                 0,927    0,226    0,822      0,927    0,871      0,713    0,897     0,860     N
Weighted Avg.    0,855    0,154    0,860      0,855    0,854      0,713    0,897     0,881

=== Confusion Matrix ===

    a    b   <-- classified as
 2252  658 |    a = S
  241 3041 |    b = N
```

Figura 14. Precisión y matriz de confusión del árbol para todos los programas
Fuente: Elaboración Propia.

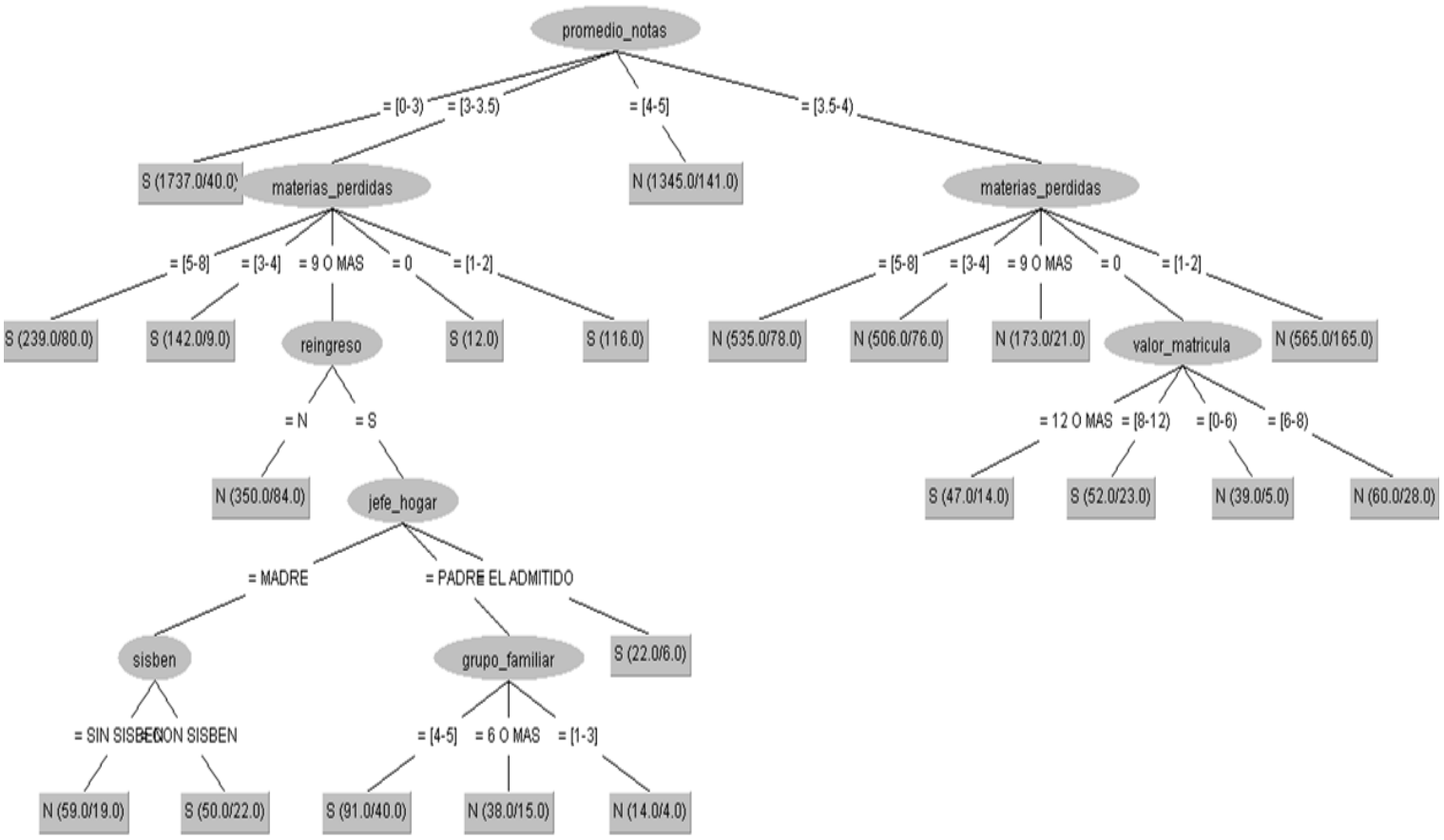

Figura 15 Árbol de decisión para todos los programas.
fuente: Elaboración Propia

3.4.3.2. Árbol de decisión para Ciencias Naturales

```
Time taken to build model: 0.03 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1977               84.4872 %
Incorrectly Classified Instances       363               15.5128 %
Kappa statistic                          0.6907
Mean absolute error                      0.2227
Root mean squared error                  0.3357
Relative absolute error                 45.3036 %
Root relative squared error             67.7061 %
Total Number of Instances             2340

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,794    0,089    0,920      0,794    0,853      0,699    0,891     0,923     S
                 0,911    0,206    0,773      0,911    0,836      0,699    0,891     0,812     N
Weighted Avg.    0,845    0,140    0,856      0,845    0,846      0,699    0,891     0,875

=== Confusion Matrix ===

    a    b   <-- classified as
 1050  272 |    a = S
   91  927 |    b = N
```

Figura 16 Precisión y matriz de confusión del árbol de clasificación para Ciencias Naturales
Fuente: Elaboración propia.

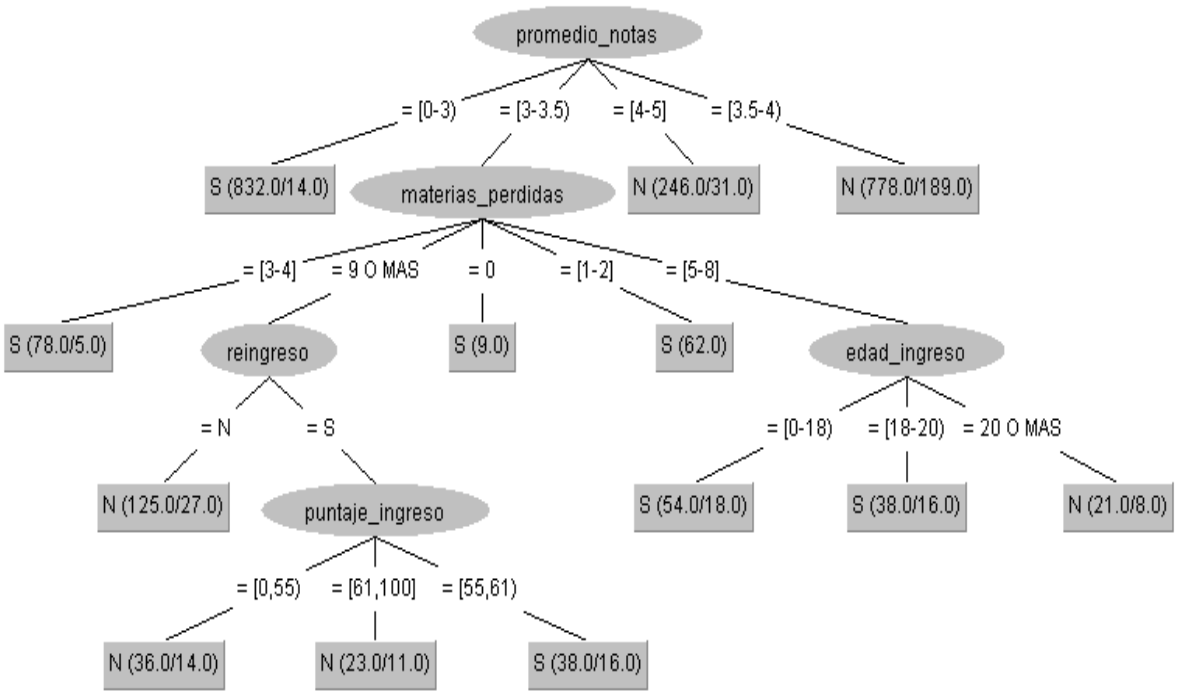

Figura 17. Árbol de decisión para programas de Ciencias Naturales.
Fuente: Elaboración propia.

3.4.3.3. Árbol de decisión para Ciencias Sociales y Humanas

```
Time taken to build model: 0.25 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        3310               85.9294 %
Incorrectly Classified Instances       542               14.0706 %
Kappa statistic                          0.7025
Mean absolute error                      0.2132
Root mean squared error                  0.3291
Relative absolute error                 43.9925 %
Root relative squared error             66.848  %
Total Number of Instances             3852

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0,749     0,063     0,893       0,749   0,814      0,710    0,889     0,884     S
               0,937     0,251     0,842       0,937   0,887      0,710    0,889     0,874     N
Weighted Avg.  0,859     0,174     0,863       0,859   0,857      0,710    0,889     0,878

=== Confusion Matrix ===

   a    b   <-- classified as
1189  399 |    a = S
 143 2121 |    b = N
```

Figura 19. Precisión y matriz de confusión del árbol de clasificación de los programas de Ciencias Sociales y Humanas. Fuente: Elaboración propia.

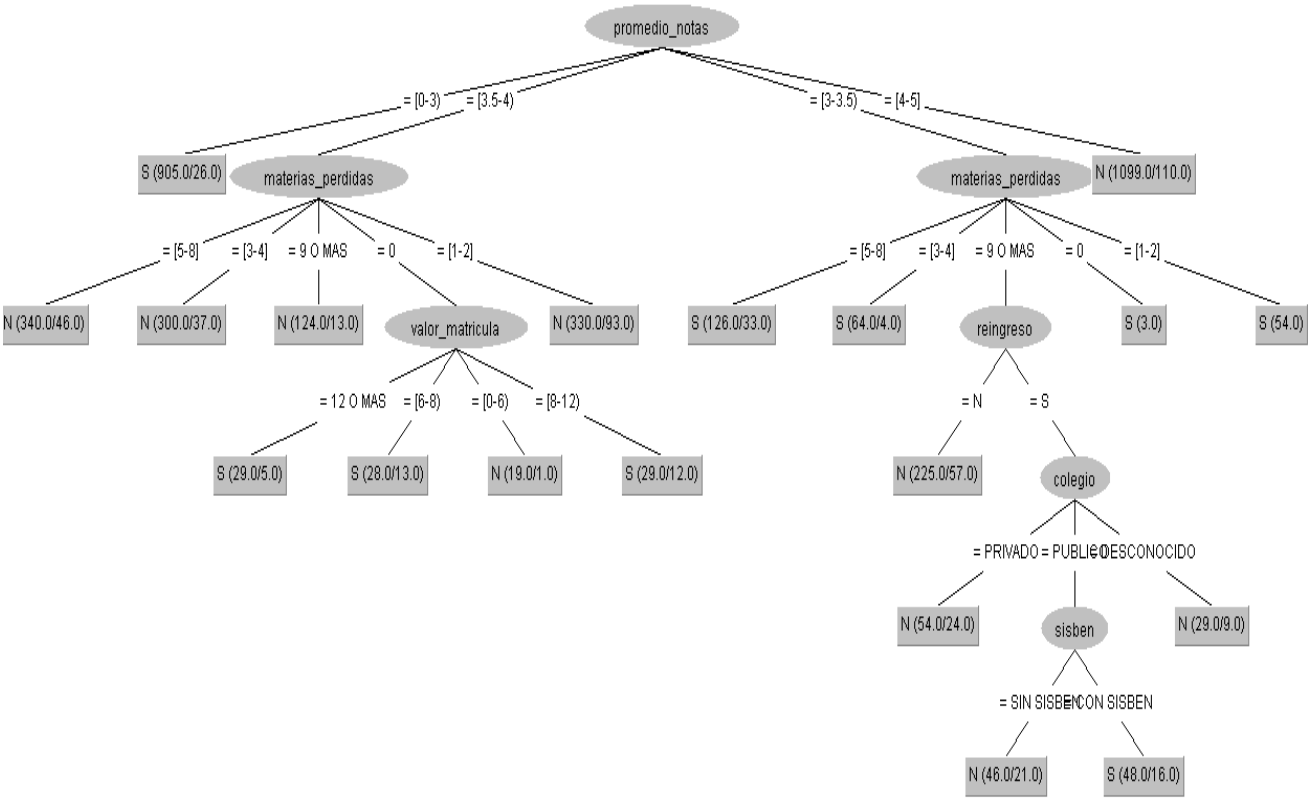

Figura 18. Árbol de decisión para programas de Ciencias Sociales y Humanas. Fuente: Elaboración propia.

4. RESULTADOS

Para escoger los patrones más representativas se tuvo en cuenta dos parámetros: el soporte que es el porcentaje mínimo de registros que deben llegar a la hoja del árbol para que la regla sea considerada y la confianza que es el porcentaje mínimo de registros que deben estar bien clasificados en la regla, el cual se calcula tomando el número total de registros que llegan a la hoja (primer número que aparece en la hoja de cada árbol) menos el número de registros mal clasificados (que corresponde al segundo número, separado por "/") y esto multiplicado por 100 y dividido entre el número total de registros que llegan a la hoja (primer número). Por ejemplo, si se genera un árbol con 2340 registros y en el árbol aparee en una hoja S(468/158). Quiere decir, que el número 468 es la cantidad total de registros que llegan a la hoja; 158 es el número de registros mal clasificados y 468-158 igual a 310 es el número de registros bien clasificados. Ahora el soporte de esta regla es del 20% (468*100/2340) y la confianza es de 66.23% (310*100/468).
Las reglas de clasificación que se consideraron fueron aquellas que superaron el 60 % de confianza y un soporte mínimo del 1%. las reglas que se analizaron se describen a continuación.

4.1. Resultados Obtenidos Considerando Todos Los Programas Y Todos Los Aspectos

La Tabla 32 muestra las clasificaciones generales que destacaron a los desertores de todos los programas de la Universidad de Nariño sede Pasto. De acuerdo al árbol de decisión de la Figura 15 estas clasificaciones son:

REGLA No.	ANTECEDENTE	DESERCION	% del total de estudiantes (6192) (soporte)	CONFIANZA %	% del total de desertores (2910)
1	Promedio_notas=[0-3)	S	28,05	97,7	58,31
2	Promedio_notas [3-3,5) & materias perdidas 3 o 4	S	2,15	93,7	4,57
3	Promedio_notas=[3-3,5) & materias perdidas 1 o 2	S	2,29	100%	3,99
4	Promedio_notas=[3,5-4) & materias perdidas= 0 & valor_matricula = 12 o mas	S	0,76	70,2	1,13

Tabla 32. Clasificación desertores para todos los programas.
Fuente: Elaboración propia.

Al igual que Rico [23] y Malagón et al. [13], el factor más influyente en la deserción de los estudiantes de la Universidad de Nariño Sede Pasto es el bajo rendimiento académico evidenciado en la Regla No. 1 (promedio de notas es menor a 3). Cabe notar que, más de la mitad de desertores (58,31%) cumplen con este patrón. En el

mismo sentido, las Reglas 2 y 3 hacen notar el predominio de los factores académicos en la deserción; tener un promedio de notas entre 3 y 3.5 y perder 1 o más materias (bajo rendimiento académico). Por su parte, la Regla No. 4 se relaciona con el factor económico, como lo es un alto valor de la matrícula (mayor a 12 smldv), con lo cual, se ratifica los resultados obtenidos por Timarán y Jiménez en [32].

4.2. Resultados Obtenidos Considerando Todos Los Aspectos Para Los Programas De Ciencias Naturales

La Tabla 33 muestra las clasificaciones que destacan a los desertores de los programas pertenecientes a Ciencias Naturales. De acuerdo al árbol de decisión de la Figura 17 estas clasificaciones son.

REGLA No.	ANTECEDENTE	DESERCION	% del total de estudiantes (2340) soporte	CONFIANZA %	% del total de desertores (1322)
1	Promedio_nota=[0-3)	S	35,55	98,31	61,87
2	Promedio_notas=[3-3,5) & materias perdidas =3 o 4	S	3.33	93,58	5,52
3	Promedio_notas[3-3,5) & materias_perdidas= 1 o 2	S	2,64	100%	4,68

Tabla 33. Clasificación desertores para programas de Ciencias Naturales.
Fuente: Elaboración propia.

Para los programas pertenecientes a las Ciencias Naturales se evidencian las mismas reglas del caso anterior. Es más, aumentan el soporte y la confianza, y por ende, acumulan una mayor cantidad de desertores. Es decir, la deserción los programas de Ciencias Naturales está ampliamente relacionada con el bajo rendimiento académico.

4.3. Resultados Obtenidos Considerando Todos Los Aspectos Para Los Programas De Ciencias Sociales Y Humanas

La Tabla 34 muestra las clasificaciones principales que de acuerdo con el árbol de decisión inciden en los desertores de los programas pertenecientes a Ciencias Sociales y Humanas.

REGLA No.	ANTECEDENTE	DESERCION	% del total de estudiantes(3852) Soporte	CONFIANZA %	% del total de desertores (1588)
1	Promedio_nota=[0-3)	S	23,5	97,12	55,35
2	Promedio_notas[3-3,5) & materias perdidas =5 a 8	S	3,27	73,80	5,85
3	Promedio_notas[3-3,5) & materias_perdidas= 3 o 4	S	1,7	93,75	3,77
4	Promedio_notas[3-3,5) & materias_perdidas= 1 o 2	S	1,40	100	3,40
5	Promedio_notas[3,5-4) & materias_perdidas= 0 & valor_matricula 12 o mas	S	0,75	82,75	1,51

Tabla 34. Clasificación desertores programas Ciencias Sociales y Humanas.
Fuente: Elaboración propia.

Se observa que los desertores de los programas pertenecientes a las Ciencias Sociales y Humanas, también están influenciados por los factores ya mencionados.

Ratificando nuevamente, que el bajo rendimiento académico es la principal causa de deserción en la Universidad de Nariño Sede Pasto, seguida del alto costo en el valor de la matrícula. Además, los resultados muestran que no hay diferencias significativas entre los factores de deserción tanto para las Ciencias Naturales como para las Ciencias Humanas.

4.4. Resultados Obtenidos Considerando Atributos Académicos

Ahora bien, si se considera únicamente aspectos académicos e institucionales, se reafirman las reglas anteriores, como se puede ver en la Tabla 35, sin embargo, para los programas pertenecientes a Ciencias Sociales y Humanas (Tabla 37) aparece un nuevo atributo (alto puntaje de ingreso), que si bien son pocos estudiantes causa incertidumbre, puesto que al tener un buen rendimiento académico (sin materias perdidas y promedio alto), genera la hipótesis: “se cambiaron de carrera”.

ANTECEDENTE	DESERCION	% del total de estudiantes(6192) soporte	CONFIANZA %	% del total de desertores (2910)
Promedio_nota=[0-3)	S	21,59	97,69	58,31
Promedio_notas=[3-3,5) & materias perdidas =3 o 4	S	2,29	93,66	4,57
Promedio_notas[3-3,5) & materias_perdidas= 1 o 2	S	1,8	100%	3,98

Tabla 35. Clasificación desertores con aspectos académicos e institucionales para todos los programas.
Fuente: Elaboración propia.

ANTECEDENTE	DESERCION	% del total de estudiantes(2340) soporte	CONFIANZA %	% del total de desertores (1322)
Promedio_nota=[0-3)	S	35,55	98,32	61,87
Promedio_notas=[3-3,5) & materias perdidas =3 o 4	S	3,33	93,59	5,52
Promedio_notas[3-3,5) & materias_perdidas= 1 o 2	S	2,6	100%	4,68

Tabla 36. Clasificación desertores con aspectos académicos e institucionales para los programas de Ciencias Naturales. Fuente: Elaboración propia.

ANTECEDENTE	DESERCION	% del total de estudiantes(3852) Soporte	CONFIANZA %	% del total de desertores (1588)
Promedio_nota=[0-3)	S	23,49	97,12	55,35
Promedio_notas=[3-3,5) & materias perdidas =5 a 8	S	3,27	73,80	5,85
Promedio_notas[3-3,5) & materias_perdidas= 3 o 4	S	1,66	93,75	3,77
Promedio_notas[3-3,5) & materias_perdidas= 1 o 2	S	1,40	100	3,40
Promedio_notas[3,5-4) & materias_perdidas= 0 & puntage_ingreso = [61-100]	S	0,67	76,92	1,25

Tabla 37. Clasificación desertores con aspectos académicos e institucionales para los programas de Ciencias Sociales y Humanas.
Fuente: Elaboración propia.

4.5. Resultados Obtenidos Considerando Atributos Socioeconómicos

Dado que los factores académicos e institucionales presentan mayor influencia en el hecho de ser o no desertor, se decide analizar la clasificación usando sólo atributos socioeconómicos, en consecuencia, se genera el árbol con atributos socioeconómicos para todos los programas, donde, se puede notar que, al disminuir el porcentaje de bien clasificados se pierde confiabilidad en los patrones que determinan la deserción, sin embargo, en la Tabla 38 se rescatan las siguientes reglas de decisión:

ANTECEDENTE	DESERCION	% del total de estudiantes(6192) soporte	CONFIANZA %	% del total de desertores (2910)
Valor_matricula=[6-8)& sexo=M	S	18,71	58,06	23,12
Valor_matricula=[8-12)& sexo=M	S	17,40	57,97	21,47
Valor_matricula=[12 o más)& sexo=M	S	13,95	58,10	17,25
Valor_matricula=[6-8)& sexo=F residencia= arrendada o anticresada & SISBEN= con SISBEN	S	2,90	58,33	3,60
Valor_matricula=[12 o más)& sexo=F& edad_ingreso= 20 o mas	S	1,64	55,88	1,95

Tabla 38. Clasificación desertores con aspectos socioeconómicos para todos los programas.
Fuente: Elaboración propia.

Aparece como factor de deserción el ya mencionado alto valor de matrícula, acompañado del hecho de ser de género masculino. Pero, genera sospecha a los investigadores, pues no se tienen un soporte y una confianza convincentes, por lo cual, hablar de género como factor de deserción en la Universidad de Nariño no sería muy adecuado.

Con el ánimo de diferenciar aspectos socioeconómicos entre los programas de Ciencias Naturales y de Ciencias Sociales y Humanas se realiza la clasificación de desertores para cada uno de los grupos, pero, nuevamente las reglas de deserción son muy similares, como lo muestran las siguientes tablas.

ANTECEDENTE	DESERCION	% del total de estudiantes(2340) soporte	CONFIANZA %	% del total de desertores (1322)
Valor_matricula=[8-12)	s	28,37	64,30	32,29
Valor_matricula=[6-8)& sexo=M	s	22,86	63,36	25,64
Valor_matricula=[12 o más)	s	20	66,23	23,44
Valor_matricula [6-8), sexo=F, estrato=alto, comuna >8	s	1,02	70,83	1,28

Tabla 39. Clasificación desertores con aspectos socioeconómicos para los programas de Ciencias Naturales.
Fuente: Elaboración propia.

ANTECEDENTE	DESERCION	% del total de estudiantes (3852) soporte	CONFIANZA %	% del total de desertores (1588)
Valor_matricula=[12 o más)& sexo=M & grupo_familiar=[1-3]	S	4,80	61,08	7,11
Valor_matricula=[6-8)& ingreso_familiar=10 a 15 & SISBEN= con SISBEN& residencia = arrendada o anticresada	S	1,81	65,71	2,89
Valor_matricula=[8-12) sexo=M estrato=rural	S	1,73	64,17	2,70
Valor_matricula=[6-8)& ingreso_familiar=[0-5)&estrato=bajo&vive_familia=N	S	1,56	66,66	2,51
Valor_matricula=[8-12)& sexo=M& estrato=bajo&jefe_hogar=madre & grupo familiar=4 o 5	S	0,87	61,11	2,07
Valor_matricula=[8-12)& sexo=M& estrato=bajo& jefe_hogar=padre & vive familia =N	S	1,03	62,5	1,57

Tabla 40. Clasificación desertores con aspectos socioeconómicos para los programas de Ciencias Sociales y Humanas. Fuente: Elaboración propia.

En general, en cuanto a los aspectos socioeconómicos se puede decir que la principal razón de deserción universitaria es el dinero, evidenciado principalmente en el alto valor de la matrícula, aunque aparecen a tributos como: ingresos familiares anuales menores a 5 salarios mínimos mensuales, pertenecer a estrato bajo, ser del sector rural y vivir en residencia arrendada o anticresada. Lo social, no parece tener mucho peso, sin embargo, el aspecto género es considerado en la mayoría de dichas reglas.

5. CONCLUSIONES

La técnica de minería de datos utilizada para la determinación de los factores asociados a la deserción en la Universidad de Nariño, fue la clasificación mediante árboles de decisión bajo el método j48, quien en primera instancia fue el que mejor exactitud mostró en comparación a los modelos Bayesnet, Random Forest, y LMT, además, j48 mostró mayor precisión y sensibilidad en sus resultados, situación advertida en [12] [17]y [25]

Las matrices de confusión de los árboles generados por j48, permiten evidenciar que existe siempre una baja tasa de falsos positivos (estudiantes que no son desertores y que son considerados desertores), lo cual, muestra que, si se crean programas de apoyo dirigidos únicamente a estudiantes considerados como posibles desertores, se acertaría en casi toda la población escogida. Además, cabe notar, que, si en la Universidad de Nariño se llevan a cabo programas orientados a retención estudiantil, sólo con estudiantes calificados como futuros desertores, se estaría incluyendo alrededor del 80% de los desertores, y, por ende, tan sólo un 20% quedarían por fuera de los proyectos (falsos negativos). De ahí, que los modelos obtenidos son de gran ayuda para para la toma de decisiones por parte de las directivas de la Universidad de Nariño, si deciden invertir en programas de retención estudiantil.

En los programas académicos considerados dentro de las Ciencias Sociales y Humanas se evidenció un mayor porcentaje de confianza en los factores académicos e institucionales en comparación con los factores socioeconómicos. Situación parecida se vivió para los programas considerados como Ciencias Naturales. Con lo cual se muestra, que los aspectos académicos fueron los más influyentes en la deserción de los estudiantes de la Universidad de Nariño.

En Ciencias Sociales y Humanas, la regla que identifica a los desertores, en cuanto a los factores académicos e institucionales es tener un promedio de notas menor a 3. Mientras, que, para los factores socioeconómicos, la regla primordial es pagar un valor de matrícula de 12 o más salarios mínimos diarios, ser de sexo masculino y tener un grupo familiar conformado por 1, 2 o 3 personas. En Ciencias Naturales, con respecto a los factores académicos e institucionales, la regla es la misma. Y con relación a los factores socioeconómicos, la regla es pagar un valor de matrícula entre 8 y menos de los 12 salarios mínimos legales diarios vigentes. Es decir, no hay diferencias significativas en los patrones de deserción obtenidos tanto para los programas de Ciencias Naturales como de Ciencias Sociales y Humanas (bajo rendimiento y alto costo de la matrícula).

En general, se evidencia con un 58,31% de representación en sus desertores que la regla determinante de la deserción con respecto a los factores académicos de las cohortes 2008A a 2011B es tener un promedio de notas bajo (menor a 3). Al igual que el trabajo desarrollado por Rico en [23], en la Universidad Nacional de Colombia

Sede Medellín y Malagón et al., en [13] para la Universidad de los Llanos, la presente investigación reveló que el factor de mayor incidencia en la deserción son la mortalidad académica y el bajo rendimiento académico evidenciado en el promedio de notas y la repitencia de asignaturas. Ratificando los resultados de Timarán y Jiménez en [32] donde, para los Programas de pregrado de la Universidad de Nariño y la Institución Universitaria CESMAG encontraron como patrón general de deserción estudiantil el promedio bajo de notas, tener materias reprobadas en los primeros semestres de la carrera y un puntaje promedio bajo en las pruebas de estado (a diferencia, este último aspecto no fue relevante para la actual investigación).

En relación a los factores socioeconómicos, con un 23,12% de representación en sus desertores la regla que identifica a los estudiantes de las cohortes 2008A a 2011B de la Universidad de Nariño sede Pasto es pagar un valor de matrícula de 6 o más salarios mínimos diarios y ser de género masculino, lo cual, es un aspecto a tener en cuenta en Colombia, puesto que otros investigadores como Rico en [23], muestran que aunque en menor medida, también la deserción estudiantil está relacionada con causas socioeconómicas. Situación parecida para Timarán y Jiménez en [30], donde revelan que, el pagar una matrícula promedio alta (mayor que $381504) es causa de deserción. Salcedo en [26] refuerza esta afirmación, para él, las causas que determinan la deserción se pueden atribuir a varios problemas externos e internos a la universidad, problemas intrínsecos al estudiante y a otras causas.

Pertenecer a un programa acreditado conllevo a concluir que no es un factor influyente que soporte la disminución de la deserción, en muestra de ello, de los 2673 estudiantes inscritos en programas acreditados el 51.07% desertó, el cual superó el índice general de deserción presente en las cohortes 2008A a 2011B que fue del 47%.

En los programas de las Ciencias Naturales fue donde se presentó mayor índice de deserción (el 56,50% desertó) en comparación con los de las Ciencias Sociales y Humanas, además, cabe anotar que el 73.09% de los estudiantes desertores lo hizo en los primeros semestres de la carrera.

En vista de los resultados obtenidos, y con el ánimo de obtener mayor información, en cuanto a deserción estudiantil, se plantea como trabajos futuros, utilizar otras técnicas de minería de datos, tales como clustering, patrones secuenciales o reglas de asociación que permitan identificar y relacionar atributos que asocian la deserción estudiantil en la Universidad de Nariño sede Pasto.

Dado que para el desarrollo de la investigación, la mala calidad de los datos de las cohortes 2006A, 2006B 2007A y 2007B conllevó a reducir el número de cohortes a estudiar, sería prudente analizar otro grupo de estudiantes además de los involucrados, lo cual, podría arrojar resultados diferentes y quizá más acertados.

Se recomienda a las directivas de la Universidad evaluar, analizar y determinar la utilidad de los patrones obtenidos para soportar la toma de decisiones eficaces enfocadas a formular políticas y estrategias relacionadas con programas de retención estudiantil.

REFERENCIAS BIBLIOGRÁFICAS

[1] Castaño, E., Gallón, S., Gómez, K. & Vásquez, J. Análisis de los factores asociados a la deserción y graduación estudiantil universitaria. Lecturas Económicas, 65, 9 - 36. Universidad de Antioquia. Medellín, Colombia, 2006.

[2] Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. & Wirth, R. CRISP-DM 1.0: Step-by-step data mining guide. SPSS.2000.

[3] Cynthia L, Fabian G, Aplicación de Redes Bayesinas usando Weka. Universidad Tecnológica Nacional, Argentina

[4] Franco, M. Factores que influyen en el ingreso y permanencia de los estudiantes den la Universidad de la Sabana. Tesis de pregrado de Psicología, Universidad de la Sabana. Santafé de Bogotá, Colombia, 1991.

[5] Gallardo, J. Metodología para el Desarrollo de Proyectos en Minería de Datos CRISP-DM [online]. www.oldemarrodriguez.com/yahoo_site_admin/assets/docs/Documento_CRISP_DM.2385037.pdf.

[6] García, M., & Álvarez, A. (2010). Análisis de datos en WEKA–pruebas de selectividad. línea] disponible en http://www. it. uc3m. es/jvillena/irc/practicas/06-07/28. pdf. Recuperado a partir de http://www.it.uc3m.es/jvillena/irc/practicas/06-07/28.pdf.

[7] Giovagnoli, P. (2002). Determinantes de la deserción y Graduación universitaria: Una Aplicación Utilizando Modelos de Duración, Documento de Trabajo 37, Universidad Nacional de la Plata, Argentina.

[8] Hall, M., Frank, E., & Witten, I. (2011). Practical Data Mining: Tutorials. University of Waikato. Recuperado a partir de http:// www.micai.org/2012/tutorials/Weka%20tutorials%20Spanish.pdf

[9] Han, J., & Kamber, M. Data Mining: Concepts and Techniques, Third Edition (3 edition.). Burlington, MA: Morgan Kaufmann, 2001.

[10] Hernández, J., Ramírez, M. J., & Ferri, C. (2005). Introducción a la Minería de Datos. Editorial Pearson Educación SA, Madrid. Recuperado a partir de http://dspace.ucbscz.edu.bo/dspace/handle/123456789/526

[11] Hernández, J. y Ramírez, M. y Ramírez, C. Introducción a la Minería de Datos, Editorial Pearson, Madrid, España, 2004.

[12] I. Kononenko, “Inductive and Bayesian Learning in Medical Diagnosis 1 Introduction,” pp. 1–24.

[13] Malagón, L., Calderón, C. & Soto E. Estudio de la deserción estudiantil de los programas de pregrado de la Universidad de los Llanos. Villavicencio, Colombia, 2006.

[14] Ministerio de Educación Nacional. Deserción Estudiantil En La Educación Superior Colombiana. [Online]. http://www.mineducacion.gov.co/sistemasdeinformacion/1735/articles-254702_libro_desercion.pdf

[15] Ministerio de Educación Nacional-MEN. Educación Superior en Colombia. Ministerio de Educación Nacional, Bogotá, Colombia, 2007.

[16] Ministerio de Educación Nacional-MEN. Diagnóstico de la Deserción Estudiantil en Colombia. Educación Superior – Boletín Informativo. No. 7. Diciembre 2006.

[17] M. Hariz, M. Adnan, W. Husain, N. Aini, and A. Rashid, “Hybrid Approaches Using Decision Tree, Naive Bayes, Means and Euclidean Distances for Childhood Obesity Prediction,” vol. 6, no. 3, pp. 99–106, 2012.

[18] Osvaldo Spositto, Martín E. Etcheverry, Hugo L. Ryckeboer, and Julio Bossero. Aplicación de técnicas de minería de datos para la evaluación del rendimiento académico y la deserción estudiantil. International Institute of Informatics and Systemics. [Online]. http://www.iiis.org/CDs2010/CD2010CSC/CISCI_2010/PapersPdf/CA156FK.pdf

[19] Pautsh, J., La Red Martínez, D.L & Cutro, L. Minería de Datos aplicada al análisis de la deserción en la Carrera de Analista en Sistemas de Computación. [Online]. http://www.dataprix.com/files/Analisis%20de%20Desercion%20Univ_0.pdf

[20] Pautsh, J., La Red Martínez, D.L & Cutro, L. Minería de Datos aplicada al análisis de la deserción en la Carrera de Analista en Sistemas de Computación. Tesis de Grado Universidad Nacional de Misiones. Misiones, Argentina.[Online]. http://exa.exa.unne.edu.ar/depar/areas/informatica/SistemasOperativos/TFAGermanPAUTSCHFinal.pdf

[21] Vizcaino, P., Aplicación de técnicas de inducción de árboles de decisión a problemas de clasificación mediante uso de Weka. Fundación Universitaria Konrad Lorenz, Facultad de ingeniería de sistemas Bogotá, 2008

[22] Restrepo, M., López, H. & Pinzón, L. Uso de la metodología Rough Sets en un modelo de deserción académica. En memorias de *XIV Congreso Ibero Latinoamericano de investigación de operaciones CIAIO*. Bogotá, Colombia, 2008.

[23] Rico, D. Caracterización de la Deserción Estudiantil en La Universidad Nacional de Colombia Sede Medellín. Universidad Nacional de Colombia. Medellín, Colombia, 2006.

[24] Rojas, M., & González, D. (2008). Deserción estudiantil en la Universidad de Ibagué, Colombia: una lectura histórica en perspectiva cuantitativa. *Zona Próxima*.

[25] R. Lakshmy, D. Ph, D. J. P. Barker, D. Ph, S. K. D. Biswas, M. Stat, and S. Ramji, “new england journal,” pp. 865–875, 2004.

[26] Salcedo, A. Deserción universitaria en Colombia. Revista Academia y virtualidad. Universidad Militar Nueva Granada. Colombia (2010).

[27] Sánchez, J. (2014). Modelos predictivos para el estudio del abandono en centros universitarios (Tesis de Pregrado). Universidad Politécnica de Madrid, Madrid.

[28] SPADIES http://www.mineducacion.gov.co/sistemasdeinformacion/1735/w3-article-357549.html

[29] Sattler, K.-U., & Dunemann, O. (2001). SQL database primitives for decision tree classifiers. En Proceedings of the tenth international conference on Information and knowledge management (pp. 379–386). ACM. Recuperado a partir de http://dl.acm.org/citation.cfm?id=502650

[30] Timarán-Pereira, R. Detección de Patrones de Bajo Rendimiento Académico y Deserción Estudiantil con Técnicas de Minería de Datos. En memorias de *VIII Conferencia Iberoamericana en Sistemas, Cibernética e Informática*. pp. 146-150. Orlando, USA, 2009.

[31] Timarán-Pereira, S. R., Hernández-Arteaga, I., Caicedo-Zambrano, J. J., Hidalgo-Troya, A. & Alvarado-Pérez, J. C. (2016). Descubrimiento de patrones de desempeño académico con árboles de decisión en las competencias genéricas de la formación profesional. Bogotá: Ediciones Universidad Cooperativa de Colombia.

[32] Timarán-Pereira, R. & Jiménez, J. Detección de Patrones de Deserción Estudiantil en Programas de Pregrado de Instituciones de Educación Superior con CRISP-DM. En memorias del Congreso Iberoamericano de Ciencia, Tecnología, Innovación y Educación. Organización de Estados Iberoamericanos-OEI. Buenos Aires, Argentina, 2014.

[33] Tinto, V. Definir la Deserción de Perspectiva. Revista de Educación Superior N° 71, ANUIES, México, (1989).

[34] Universidad Pedagógica Nacional. La deserción estudiantil: reto investigativo y estratégico asumido de forma integral por la UPN. Tunja, Colombia, 2004.

[35] Usama, F., Piatestsky-Shapiro, G., & Smiyth, P. (1996). The kdd process for extracting useful knowledge from volumes of data. *Commun*. 39(11), 27-34.

[36] Valero Orea, S. Aplicación de técnicas de minería de datos para predecir la deserción. Universidad Tecnológica de Izúcar de Matamoros. Izúcar de Matamoros, México, 2009. http://www.utim.edu.mx/~svalero/docs/MineriaDesercion.pdf

[37] Valero, S., Salvador, A. & García, M. Minería de datos: predicción de la deserción escolar mediante el algoritmo de árboles de decisión y el algoritmo de los k vecinos más cercanos. Universidad Tecnológica de Izúcar de Matamoros. México [Online]. http://www.utim.edu.mx/~svalero/docs/e1.pdf

[38] Vallejos, S. Minería de Datos. Trabajo de Adscripción. Corrientes, Argentina. Universidad Nacional del Nordeste. Facultad de Ciencias Exactas, Naturales y Agrimensura. Licenciatura en Sistemas de Información, (2006), 31p.

[39] Zuleta, A., Chaves, A. & Rosero. La deserción de estudiantes en el programa de Ingeniería de Sistemas de la Universidad de Nariño: 2005 - 2008. Tesis de

Maestría en Educación. Facultad de Educación, Universidad de Nariño. Pasto, Colombia, 2005.

ANEXOS

Anexo A. Diccionario De Datos

Descripción Base De Datos ESTUDIANTES	
Variable	**Descripción**
Código	Contiene el código del estudiante que lo identifica dentro de la Universidad
tipo_ingreso	Tipo de ingreso, (N) normal
tipo_documento	Contiene los distintos tipo de documento que identifican al estudiante, CC, TI, CE, Pasaporte, etc
identificacion_new	Número de identificación del alumno actualizada
Ciudadnac	Ciudad donde nació el estudiante
departamentonac	Departamento donde nació el estudiante
Paisnac	País de nacimiento.
Nombres	Nombres del estudiante
Apellidos	Apellidos del estudiante
Sexo	Género que identifican al estudiante
estado_civil	Estado civil, si es soltero, casado, viudo, etc
fecha_nacimiento	Fecha de nacimiento
libreta_militar	Número de libreta militar
Distrito	Número de distrito donde se expide la libreta militar
cod_sisben	Categoría a la cual pertenece de acuerdo a SISBEN
cod_eps	Código de la empresa prestadora de salud
e_mail	Dirección de correo electrónico
dir_pasto	Dirección donde vive en Pasto
bar_pasto	Barrio donde vive en Pasto
dir_permanente	Dirección de residencia del alumno
bar_permanente	Barrio de residencia del alumno
cod_ciudadres	Código de ciudad de residencia del estudiante
cod_departamentores	Código de departamento de residencia del estudiante
cod_paisres	Código de país de residencia del estudiante
vive_familia	Si el estudiante vive con la familia en el momento de la inscripción
Estrato	Estrato donde reside el estudiante.
estrato_vivira	Estrato donde vivirá el estudiante en Pasto
tipo_residencia	Tipo de residencia: arrendada, propia, otra

etnia_code	Código de la etnia a la que pertenece
Especial	Tipo de cupo de ingreso.
cod_capacidad	Código capacidad que puede tener un estudiante. Ej.: superdotado
cod_discapacidad	Código de discapacidad que tiene un estudiante Ej.: Ceguera
necesidades_educativas	Si el estudiante presenta necesidades educativas especiales.
actualmente_trabaja	Si en el momento de la inscripción el estudiante se encuentra trabajando
tipo_trabajo	Tipo de trabajo Ej.: medio tiempo
duracion_trabajo	Duración del trabajo en horas diarias
rango_ingresos	Rango de ingresos obtenidos del trabajo.
validacion_icfes	Contiene la información del estudiante si hizo validación del ICFES.
ident_procedencia	Identificación con la que ingresa el estudiante.
cod_colegio	Código del colegio de procedencia.
tipo_colegio	Contiene el tipo de colegio en el que estudio Ej.: privado
valor_matric_colegio	Contiene el valor de la matrícula de pago en el colegio.
ano_pago_colegio	Año de pago de la última matricula en el colegio
cod_icfestotal	Código de inscripción ante el ICFES
puntaje_ingreso	Puntaje de ingreso a la Universidad
p_g_materia1	Puntaje área de Biología
p_g_materia2	Puntaje área de Matemáticas
p_g_materia3	Puntaje área de Filosofía
p_g_materia4	Puntaje área de Física
p_g_materia5	Puntaje área de Historia
p_g_materia6	Puntaje área de Química
p_g_materia7	Puntaje área de Lenguaje
p_g_materia8	Puntaje área de Geografía
p_g_materia9	Puntaje área Ingles
p_g_materia10	Puntaje área de Sociales y Ciudadanas
p_g_materia11	Puntaje área de Lectura Critica
p_g_materia12	Puntaje área de Ciencias Naturales
p_g_materia13	Puntaje área de Razonamiento Cuantitativo
p_g_materia14	Puntaje en Competencias Ciudadanas
jefe_familia	Muestra el representante de familia
numero_grupofamiliar	Número de integrantes del grupo familiar
ingresos_familiares	Promedio de ingresos familiares

ano_ingresos	Año en el que se calcularon los ingresos familiares
numero_aportantes	Número de personas aportantes en la familia
numero_hermanos	Número de hermanos en el hogar
numero_hermanosestsuperior	Número de hermanos que están en la Universidad
nueva_matricula	Valor pagado de matricula
nuevos_servicios	Valor de nuevos servicio para la matricula del semestre
pago_contado	Valor total a pagar de contado para la matricula
cod_carrera	Identificación de carrera en la que se encuentra matriculado
cod_facultad	Identificación de facultad a la cual pertenece
nombre_facultad_la	Nombre de la facultad a la cual se encuentra inscrito
nombre_carrera_lar	Nombre de la carrera en la cual se encuentra inscrito.
extension_proviene	Extensión de la que proviene el estudiante. EJ.: extensión sede Ipiales
periodo_academico	Periodo en el cual ingresó el estudiante. Ej.: 2010 A.
Egresados	Contiene información de si el estudiante es: Egresado, Graduado o No egresado
fecha_egreso	Fecha en la cual el estudiante terminó materias
fecha_grado	Fecha en la que el estudiante adquirió su título.

Tabla 41. Descripción base de datos ESTUDIANTES.
Fuente: Elaboración propia.

Descripción Base De Datos PAGOS	
Variable	**Descripción**
cod_alumno	Código de identificación del estudiante en la Universidad
nueva_matricula	Valor de la matrícula del semestre
nuevos_servicios	Valor de nuevos servicios para la matricula del semestre
pago_contado	Valor total a pagar de contado por la matricula
semestre	Semestre para el cual paga la matricula
periodo	Periodo de pago de matrícula. EJ.: 2015A
esta_vigente	Si el estudiante está vigente

Tabla 42. Descripción base de datos PAGOS.
Fuente: Elaboración propia.

Descripción Base De Datos NOTAS	
Variable	**Descripción**
cod_alumno	Código de identificación del estudiante en la Universidad
cod_materia	Código de identificación de materia
detalle_largo	Nombre de la materia
valor	Nota de la materia
fecha_inicia	Fecha en que inicia a cursar materia
fecha_terminacion	Fecha de terminación de materia

Tabla 43. Descripción base de datos NOTAS.
Fuente: Elaboración propia.

Atributo Agregado	Descripción
Edad_ingreso	Edad de ingreso del estudiante. Edad_ingreso = periodo_ingreso – Año_Nacimiento
Promedio_notas	Promedio de notas de la carrera.
semestres_transcurridos	Número de semestres transcurridos desde su ingreso hasta que termina.
semestre_matriculado	Número de semestres en los que se matriculó.
Reingreso	Si el estudiante abandona estudios por uno o más semestres y regresa.
Desertor	Si el estudiante es desertor de acuerdo a la definición por MEN
tipo_ciencia	Clasificación a cada programa de pregrado como ciencia Natural o Social
edad_egreso	Edad del estudiante cuando egreso de la universidad
edad_grado	Edad de grado del estudiante
tipo_desertor	Si el estudiante presenta una deserción inicial (de 1er semestre a 3er semestre), media (de 4to a 7mo semestre) o final(de 8vo semestre en adelante)
Smldv	Salario mínimo legal diario vigente para pago de matrícula.
programa_acreditado	Si el programa al cual está inscrito el estudiante es acreditado de alta calidad
homologado	Si el estudiante homologo materias
zona_nacimiento	Zona de nacimiento del estudiante clasificado de acuerdo a las subregiones del departamento de Nariño: CENTRO

	SUR OCCIDENTE COSTA NORTE PASTO PUTUMAYO SUR OTRA
Colegio	Si el colegio de procedencia del estudiante es privado, público o desconocido.
régimen_salud	Sistema de salud del estudiante, subsidiado o contributivo.
Comuna	Número de comuna en la que vive el estudiante 1, 2, 3, 4, etc
Residencia	Si la residencia es propia o arrendada.
Etnia	Si el estudiante pertenece o no a una etnia
zona_procedencia	Zona de procedencia del estudiante clasificado de acuerdo a las subregiones del departamento de Nariño: CENTRO SUR OCCIDENTE COSTA NORTE PASTO PUTUMAYO SUR OTRA
Estrato	Estrato de la residencia donde vive el estudiante; 0, 1, 2, 3, 4, 5, 6, otro
cupo_especial	Si el estudiante ingreso con cupo especial. (S) si, (N) no
capacidad	Si el estudiante posee un tipo de capacidad (S) si, (N) no
discapacidad	Si el estudiante presenta algún tipo de discapacidad (S) si, (N) no
Grupo_familar	Se categoriza el número de integrantes del grupo familiar del estudiante.
Ingreso_familiar	Ingresos familiares en salarios mínimos legales vigentes diarios, en el periodo de ingreso.
Hermanos_universidad	Si el estudiante tiene o no hermanos en la universidad.

doble_desertor	Si el estudiante presenta deserción primaria y finalmente abandona (S) si, (N) no
materias_perdidas	Numero de materias perdidas
Sisben	CON SISBEN si el estudiante tiene Nivel sisben entre 1 y 3, SIN SISBEN en caso contrario.
Estrato	Estrato de la vivienda de los estudiantes clasificados en ALTO (mayor a 4), BAJO (1 y 2), MEDIO (3 y 4) y RURAL (0).
Valor_matricula	Promedio del valor de la matrícula en salarios mínimos diarios.
Tipo_materia	A qué tipo de formación pertenece cada materia: BÁSICA, PROFESIONALIZACIÓN, HUMANÍSTICA o COMPLEMENTARIA.

Tabla 44. Atributos agregados a Cohortes2008A2011B.
Fuente: Elaboración propia.

Anexo B. Conteos

FACULTAD	ESTUDIANTES	%
ARTES	2779	15,87
CIENCIAS AGRICOLAS	1266	7,23
CIENCIAS DE LA SALUD	571	3,26
CIENCIAS ECONOMICAS Y ADMINISTRATIVAS	2045	11,68
CIENCIAS EXACTAS Y NATURALES	2277	13,00
CIENCIAS HUMANAS	3226	18,42
CIENCIAS PECUARIAS	1414	8,08
DERECHO	1059	6,05
EDUCACION	1028	5,87
INGENIERIA	1364	7,79
INGENIERIA AGROINDUSTRIAL	481	2,75
TOTAL	**17510**	**100**

Tabla 45. Estudiantes por facultad.
Fuente: Elaboración propia.

PERIODO	No ESTUDIANTES
2006A	249
2006B	1416
2007A	237
2007B	1414
2008A	219
2008B	1358
2009A	193
2009B	1475
2010A	144
2010B	911
2011A	917
2011B	975
2012A	965
2012B	951
2013A	934
2013B	1046
2014A	988
2014B	1003
2015A	1156
2015B	959
TOTAL	17510

Tabla 46. Estudiantes por periodo de ingreso.
Fuente: Elaboración propia.

VARIABLE	TOTAL DE NULOS	% DE NULOS
Código	0	0,0%
tipo_ingreso	0	0,0%
tipo_documento	3316	18,9%
identificacion_new	3316	18,9%
Ciudadnac	4270	24,4%
Departamentonac	4270	24,4%
Paisnac	4271	24,4%
Nombres	0	0,0%
Apellidos	0	0,0%
Sexo	3316	18,9%
estado_civil	0	0,0%
fecha_nacimiento	0	0,0%
libreta_militar	16145	92,2%
Distrito	16288	93,0%
cod_sisben	0	0,0%
cod_eps	3316	18,9%
e_mail	4239	24,2%
dir_pasto	3406	19,5%
bar_pasto	3862	22,1%
dir_permanente	3317	18,9%
bar_permanente	3838	21,9%
cod_ciudadres	3316	18,9%
cod_departamentores	3316	18,9%
cod_paisres	3316	18,9%
vive_familia	3316	18,9%
Estrato	0	0,0%
estrato_vivira	0	0,0%
tipo_residencia	0	0,0%
etnia_code	0	0,0%
Especial	0	0,0%
cod_capacidad	0	0,0%
cod_discapacidad	0	0,0%
necesidades_educativas	3915	22,4%
actualmente_trabaja	4798	27,4%
tipo_trabajo	0	0,0%
duracion_trabajo	0	0,0%
rango_ingresos	0	0,0%
validacion_icfes	3317	18,9%
ident_procedencia	3316	18,9%
cod_colegio	0	0,0%

tipo_colegio	0	0,0%
valor_matric_colegio	0	0,0%
ano_pago_colegio	0	0,0%
cod_icfestotal	3316	18,9%
puntaje_ingreso	0	0,0%
p_g_materia1	0	0,0%
p_g_materia2	0	0,0%
p_g_materia3	0	0,0%
p_g_materia4	0	0,0%
p_g_materia5	0	0,0%
p_g_materia6	0	0,0%
p_g_materia7	0	0,0%
p_g_materia8	0	0,0%
p_g_materia9	0	0,0%
p_g_materia10	249	1,4%
p_g_materia11	0	0,0%
p_g_materia12	0	0,0%
p_g_materia13	0	0,0%
p_g_materia14	0	0,0%
jefe_familia	0	0,0%
numero_grupofamiliar	0	0,0%
ingresos_familiares	0	0,0%
ano_ingresos	0	0,0%
numero_aportantes	0	0,0%
numero_hermanos	0	0,0%
numero_hermanosestsuperior	0	0,0%
nueva_matricula	0	0,0%
nuevos_servicios	0	0,0%
pago_contado	0	0,0%
cod_carrera	0	0,0%
cod_facultad	0	0,0%
nombre_facultad_la	0	0,0%
nombre_carrera_lar	0	0,0%
extension_proviene	0	0,0%
periodo_academico	0	0,0%
Egresados	2	0,0%
fecha_egreso	12047	68,8%
fecha_grado	13400	76,5%

Tabla 47. Datos nulos, repositorio histórico.
Fuente: Elaboración Propia.

VARIABLE	% NULOS Y FALTANTES
codigo	0%
tipo_ingreso	0%
tipo_documento	100%
identificacion_new	100%
ciudadnac	100%
departamentonac	100%
paisnac	100%
nombres	100%
apellidos	100%
sexo	100%
estado_civil	100%
fecha_nacimiento	100%
libreta_militar	100%
distrito	100%
cod_sisben	100%
cod_eps	100%
e_mail	100%
dir_pasto	100%
bar_pasto	100%
dir_permanente	100%
bar_permanente	100%
cod_ciudadres	100%
cod_departamentores	100%
cod_paisres	100%
vive_familia	100%
estrato	100%
estrato_vivira	100%
tipo_residencia	100%
etnia_code	100%
especial	100%
cod_capacidad	100%
cod_discapacidad	100%
necesidades_educativas	100%
actualmente_trabaja	100%
tipo_trabajo	100%
duracion_trabajo	100%
rango_ingresos	100%
validacion_icfes	100%
ident_procedencia	100%
cod_colegio	100%

tipo_colegio	100%
valor_matric_colegio	100%
ano_pago_colegio	100%
cod_icfestotal	100%
puntaje_ingreso	0%
p_g_materia1	0%
p_g_materia2	0%
p_g_materia3	0%
p_g_materia4	0%
p_g_materia5	0%
p_g_materia6	0%
p_g_materia7	0%
p_g_materia8	0%
p_g_materia9	0%
p_g_materia10	8%
p_g_materia11	100%
p_g_materia12	100%
p_g_materia13	100%
p_g_materia14	100%
jefe_familia	100%
numero_grupofamiliar	100%
ingresos_familiares	100%
ano_ingresos	100%
numero_aportantes	100%
numero_hermanos	100%
numero_hermanosestsuperior	100%
nueva_matricula	0%
nuevos_servicios	0%
pago_contado	0%
cod_carrera	0%
cod_facultad	0%
nombre_facultad_la	0%
nombre_carrera_lar	0%
extension_proviene	0%
periodo_academico	0%
egresados	0%
fecha_egreso	51%
fecha_grado	55%

Tabla 48. Nulos y faltantes, periodos 2006A a 2007B.
Fuente: Elaboración propia.

Anexo C. Categorizaciones

No	ATRIBUTO	DESCRIPCION	NÚMERO DE CATEGORÍAS
1	nombre_facultad_la	{ARTES, CIENCIAS AGRICOLAS, …, INGENIERIA AGROINDUSTRIAL}	11
2	nombre_carrera_lar	{ADMINISTRACION DE EMPRESAS, ARQUITECTURA, …, ZOOTECNIA}	37
3	Reingreso	{S,N}	2
4	Desertor	{S,N}	2
5	tipo_ciencia	{NATURALES, SOCIALES Y HUMANAS}	2
6	programa_acreditado	{S,N}	2
7	puntaje_ingreso	VARIABLE CATEGORIZADA EN INTERVALOS	3
8	promedio_notas	VARIABLE CATEGORIZADA EN INTERVALOS	4
9	Colegio	{DESCONOCIDO,PUBLICO,PRIVADO}	3
10	edad_ingreso	VARIABLE CATEGORIZADA EN INTERVALOS	3
11	materias perdidas	VARIABLE CATEGORIZADA EN INTERVALOS	5

Tabla 49. Atributos académicos e institucionales cohortes 2008A2011B.
Fuente: Elaboración propia.

No	ATRIBUTO	DESCRIPCION	NÚMERO DE CATEGORÍAS
1	Sexo	{F,M}	2
2	vive_familia	{S,N}	2
3	zona_procedencia	{ CENTRO, SUR OCCIDENTE, COSTA, NORTE, OTRA, PASTO, PUTUMAYO, SUR }	8
4	régimen_salud	{CONTRIBUTIVO, SUBSIDIADO}	2
5	Comuna	{0,1,2,3,…,11,12}	13
6	Estrato	{ALTO, BAJO, MEDIO, RURAL}	4
7	Sisben	{CON SISBEN, SIN SIBEN}	2
8	grupo_familiar	Variable categorizada en intervalos	3
9	ingreso_familiar	Variable categorizada en intervalos	4
10	hermanos_universidad	{S,N}	2
11	estado_civil	{CASADO, DIVORCIADO, NO INFORMA, SOLTERO, UNION LIBRE, VIUDO}	6
12	Edad_ingreso	Variable categorizada en intervalos	3
13	Residencia	{ARRENDADA O ANTICRESADA, PROPIA}	2
14	jefe_hogar	{EL ADMITIDO, MADRE, PADRE}	3
15	valor_matricula	Variable categorizada en intervalos	4
16	Desertor	{S,N}	2

Tabla 50. Atributos socioeconómicos cohortes 2008A2011B.
Fuente: Elaboración propia.

PUNTAJE_INGRESO	
Categorización	**No Estudiantes**
[0,55)	2085
[55,61)	2162
[61,100]	1945

Tabla 51. Categorización variable PUNTAJE_INGRESO.
Fuente: Elaboración propia.

GRUPO_FAMILIAR	
Categorización	**No Estudiantes**
[1-3]	1652
[4-5]	3396
6 O MÁS	1144

Tabla 52. Categorización variable GRUPO_FAMILIAR.
Fuente: Elaboración propia.

INGRESO_FAMILIAR	
Categorización	**No Estudiantes**
[0-5)	1521
[10-15)	1754
[5-10)	1363
15 O MÁS	1554

Tabla 53.Categorización variable INGRESO_FAMILIAR en SMLV Mensuales.
Fuente: Elaboración propia.

PROMEDIO_NOTAS	
Categorización	**No Estudiantes**
[0-3)	1737
[3-3.5)	1133
[3.5-4)	1977
[4-5]	1345

Tabla 54. Categorización variable PROMEDIO_NOTAS.
Fuente: Elaboración propia.

EDAD_INGRESO	
Categorización	**No Estudiantes**
[0-18)	2270
[18-20)	2120
20 O MÁS	1802

Tabla 55. Categorización variable EDAD_INGRESO.
Fuente: Elaboración propia.

VALOR_MATRICULA	
Categorización	**No Estudiantes**
[0-6)	1133
[6-8)	1883
[8-12)	1736
12 O MÁS	1440

Tabla 56. Categorización variable VALOR_MATRICULA en SMLV Diarios.
Fuente: Elaboración propia.

MATERIAS_PERDIDAS	
Categorización	**No Estudiantes**
0	960
[1-2]	1241
[3-4]	1160
[5-8]	1653
9 O MÁS	1178

Tabla 57. Categorización variable MATERIAS_PERDIDAS.
Fuente: Elaboración propia.

Anexo D. Clasificaciones

SUBREGIONES DE NARIÑO	MUNICIPIOS
CENTRO	ANCUYA, BUESACO, CHACHAÜÍ, CONSACÁ, EL PEÑOL, EL TAMBO, FUNES, LA FLORIDA, NARIÑO, SANDONÁ, TANGUA,YACUANQUER
COSTA	BARBACOAS, EL CHARCO, FRANCISCO PIZARRO, LA TOLA, MAGÜÍ PAYÁN, MOSQUERA, OLAYA HERRERA, ROBERTO PAYÁN, SANTA BARBARA, TUMACO
NORTE	ÁLBAN,ARBOLEDA, BELÉN, COLÓN, CUMBITARA, EL ROSARIO, EL TABLÓN DE GOMEZ, LA CRUZ, LA UNIÓN, LEIVA POLICARPA, SAN BERNARDO, SAN LORENZO, SAN PABLO, SAN PEDRO DE CARTAGO,TAMINANGO.
PASTO	SAN JUÁN DE PASTO
SUR	ALDANA, CONTADERO, CÓRDOBA, CUASPUD, CUMBAL, GUACHUCAL GUALMATÁN, ILES, IPIALES, POTOSÍ PUERRES, PUPIALES
SUROCCIDENTE	GUAITARILLA, IMUÉS, LA LLANADA, LINARES, LOS ANDES, MALLAMA, OSPINA, PROVIDENCIA, RICAURTE SAMANIEGO, SANTACRUZ, SAPUYES TÚQUERRES

Tabla 58. Clasificación por Zonas para el departamento de Nariño.
Fuente: https://es.wikipedia.org/wiki/Nari%C3%B1o_(Colombia)

Anexo E. Limpieza

VARIABLE	DESCRIPCION DE LIMPIEZA
Ciudadnac	Adición y corrección de ciudades
Departamentonac	Adición y corrección de departamentos
Paisnac	Adición y corrección de países
Nombres	Actualización a NA (no aplica)
Apellidos	Actualización a NA (no aplica)
Sexo	Actualización a F y M
estado_civil	Identificación de cada tipo de estado civil 1 = soltero, 2 = casado 3= divorciado, 4 = viudo, 5 = Unión libre 6 = desconocido
fecha_nacimiento	Corrección de fechas de nacimiento y adición de NA
cod_sisben	Identificación a que categoría del sisben pertenece.
cod_eps	Identificación de códigos EPS.
e_mail	Actualización NA (no aplica)
bar_pasto	Corrección del nombre de barrio
cod_ciudadres	Corrección de códigos de ciudades
cod_departamentores	Corrección de códigos de departamento
cod_paisres	Corrección de códigos de países
Estrato	Corrección de los estratos actualización de 0 a 6 y OTRO
estrato_vivira	Actualización de 0 a 6 y como desconocido y OTRO
tipo_residencia	Corrección de nombres
validacion_icfes	Actualización a SI, NO
tipo_colegio	Actualización a (público, privado o desconocido)
puntaje_ingreso	Se corrigió ponderados de ingreso fuera de rango.
p_g_materia1	Mediante la utilización del documento de identidad y código de registro ante el ICFES se corrigió el nombre de cada materia Biología, matemáticas, filosofía, física, historia, química lenguaje, geografía, inglés, sociales y
p_g_materia2	
p_g_materia3	
p_g_materia4	
p_g_materia5	
p_g_materia6	
p_g_materia7	
p_g_materia8	

p_g_materia9	ciudadanas, lectura crítica, ciencias naturales, razonamiento cuantitativo, competencias ciudadanas.
p_g_materia10	
p_g_materia11	
p_g_materia12	
p_g_materia13	
p_g_materia14	
jefe_familia	Adición de tipo desconocido
numero_grupofamiliar	Adición de tipo desconocido
ingresos_familiares	Corrección de datos incoherentes, adición de tipo desconocido
ano_ingresos	Corrección de años
numero_aportantes	Adición de tipo desconocido
cod_carrera	Corrección de códigos que identifican la carrera dentro de la Universidad
nombre_facultad_la	Aclaración de nombres de la facultad a la cual pertenece
nombre_carrera_lar	Actualización del nombre de la carrera.
fecha_inicia	Corrección de fechas en que inicia a cursar materia
fecha_terminacion	Corrección de fecha de terminación de materia

Tabla 59. Descripción limpieza de variables.
Fuente: Elaboración propia.

Atributo eliminado	Razón
Código	Variable identificadora
tipo_ingreso	Todos poseen N, sin ningún valor para el proceso de minería
tipo_documento	Son variables identificadoras que no aportan al proceso de minería
identificacion_new	
Ciudadnac	Correlacionadas con zona_nacimiento
Departamentonac	
Paisnac	
Nombres	Variables identificadoras sin aporte al proceso de minería
Apellidos	
fecha_nacimiento	Correlacionada con los atributos edad_ingreso, edad_egreso, edad_grado
Edad_egreso	Alto porcentaje de NA (muchos estudiantes que aún no egresan).
Edad_grado	
libreta_militar	Variable identificadora y además las mujeres no poseen este documento y alto porcentaje de nulos.
Distrito	Variable identificadora sin aporte para el proceso de minería, alto porcentaje de nulos.
cod_sisben	Correlacionada con sisben
cod_eps	Correlacionada con régimen_salud
e_mail	Variable identificadora sin aporte al proceso de minería, alto porcentaje de nulos
dir_pasto	Correlacionada con comuna
bar_pasto	
dir_permanente	Correlacionada Zona_procedencia
bar_permanente	
cod_ciudadres	
cod_departamentores	
cod_paisres	
estrato_vivira	Correlacionada con comuna.
tipo_residencia	Correlacionado con residencia.

etnia_code	Sin aporte al proceso de minería.
Especial	Cambio a cupo_especial de tipo (SI, NO)
cod_capacidad	Cambio a capacidad de tipo (SI, NO)
cod_discapacidad	Poca importancia para el proceso de minería
tipo_trabajo	Correlacionado con actualmente_trabaja
duracion_trabajo	
rango_ingresos	
actualmente_trabaja	Sin aporte al proceso de minería (muy pocos trabajan)
validacion_icfes	Presenta información incoherente
ident_procedencia	Variable identificadora
cod_colegio	Cambio a colegio
tipo_colegio	Presenta información incoherente
valor_matric_colegio	En colegios públicos no se paga matricula
ano_pago_colegio	
cod_icfestotal	Variable identificadora, sin aporte al proceso de minería
p_g_materia1	Correlacionadas con puntaje_ ingreso
p_g_materia2	
p_g_materia3	
p_g_materia4	
p_g_materia5	
p_g_materia6	
p_g_materia7	
p_g_materia8	
p_g_materia9	
p_g_materia10	
p_g_materia11	
p_g_materia12	
p_g_materia13	
p_g_materia14	
jefe_familia	Cambio a jefe_hogar
numero_grupofamiliar	Cambio a grupo_familiar
ingresos_familiares	Cambio a ingreso_familiar
ano_ingresos	Correlacionada con ingreso_familiar

numero_aportantes	Correlacionada con ingreso_familiar y grupo_familiar
numero_hermanos	Correlacionada con grupo familiar
numero_hermanosestsuperior	Cambio a hermanos_universidad
nueva_matricula	Correlacionada con valor_matricula
nuevos_servicios	
pago_contado	
cod_carrera	Variable numérica que identifica la carrera en la universidad
cod_facultad	Variable identificadora
extension_proviene	Variable donde todos sus valores son Pasto
fecha_egreso	Correlacionada con edad de egreso
fecha_grado	Correlacionada con edad de grado
Promedio	Variable de uso temporal
Semestres_transcurridos	
Semestres_matriculado	
Numerorepeticion	Variable de uso temporal
Semestreactual	Variable de uso temporal
Smldv	Variable de uso temporal
Tipo_desertor	Usadas solo para datos estadísticos, correlacionada con desertor.
Homologado	Muy pocos homologan materias
Doble_desertor	No interesa para el proceso de minería. Usada para datos estadísticos.
Fechaingreso	Correlacionada con edad_ingreso

Tabla 60. Atributos eliminados Cohortes2008A2011B.
Fuente: Elaboración propia.